



3 1761 11701349 0

CAI  
TB  
-1974  
P55 v.1

Government  
Publications

Treasury  
Board

Conseil  
du Trésor

# POLSIM

A MICRO-SIMULATION MODEL  
FOR POLICY ANALYSIS

## Volume 1

MAIN TEXT





*Presented to the*  
**LIBRARY of the**  
**UNIVERSITY OF TORONTO**  
*by*  
**D.G. HARTLE**



REPORT  
OF THE  
POLSIM PROJECT

Colin J. Hindle  
Charles B. Marriott  
George D. Mistriotis  
Planning Branch  
Treasury Board Secretariat  
Ottawa, May 1974





## CONTENTS

	<u>Preface</u>	<u>Page</u>
1.	<u>Introduction</u>	1
1.1	Motivation for the Development of POLSIM	
1.2	Microdata Simulation	
1.3	Overview of the Model	
1.4	Running the Model	
1.5	Examples of Simulation Experiments Feasible Using POLSIM	
2.	<u>Initial Year State Descriptions</u>	13
2.1	The Individual State Vector	
2.2	Alternative Sources of Initial Year Data	
2.2.1	Department of National Revenue (Taxation) Tax Analysis Data Base	
2.2.2	Unemployment Insurance Commission Data Base	
2.2.3	The Census	
2.2.4	Survey of Consumer Finance	
2.3	The Initial Model Population	
3.	<u>The Immigration Block</u>	25
3.1	Purpose and Overview	
3.2	General Structure	
3.3	Calculation of the State Vectors	
3.4	Parameter Estimation	
3.4.1	Demographic Variables	
3.4.2	The Assignment of Children	
3.4.3	Participation Rate for Married Women	
3.4.4	Employment Income Distributions	
3.4.5	Education Distributions	
3.5	Validation of the Immigration Block	



Digitized by the Internet Archive  
in 2023 with funding from  
University of Toronto



4.1 The Demographic Block Model

4.1.1 Introduction

4.1.2 Structure of the Demographic Block

4.1.3 Demographic Block Processes

a) Death

b) Emigration

c) Birth

d) Divorce

e) Marriage

f) Family Independency Process

g) Internal Migration

4.2 Demographic Block Parameters

4.2.1 Death Process Parameters

4.2.2 Emigration Process Parameters

4.2.3 Birth Process Parameters

4.2.4 Divorce Process Parameters

4.2.5 Marriage Process I Parameters

4.2.6 Marriage Process II Parameters

4.2.7 Family Independence Parameters

4.2.8 Interprovincial Migration Parameters

4.3 Validation of the Demographic Block

4.3.1 Additivity of Errors Principle

4.3.2 Simulation Errors

4.3.3 Initial Population Errors

4.3.4 Validation Results

5.1 General Overview

5.1.1 Purpose of the Activity Block

5.1.2 Methodology and Data

5.1.3 The Activity Variables

5.1.4 General Organization of the Activity Block







- 5.2 Detailed Organization and Assumptions of the Activity Block
- 5.3 The Labour Force Model
  - 5.3.1 The Basic Model
  - 5.3.2 Model Adjustment to Account for Labour Market Structural Shift
  - 5.3.3 Adjustment for Class of Person
  - 5.3.4 The Disaggregation of Transition Matrices
- 5.4 Validation of the Activity Block
  - 5.4.1 Theoretical Aspects
  - 5.4.2 Validation Results
- 5.5 Data and Inputs
  - 5.5.1 School Transitions
  - 5.5.2 Activity Transitions
  - 5.5.3 Adjustment Parameters
  - 5.5.4 Variable Parameters

## 6. The Market Income Block

155

- 6.1 The Market Income Model
  - 6.1.1 Introduction
  - 6.1.2 Overview of the Model
  - 6.1.3 The Income Processes
- 6.2 The Market Income Block Parameters
  - 6.2.1 Employment Income Parameters
  - 6.2.2 Property Income Parameters
  - 6.2.3 Retirement Income Parameters
- 6.3 Validation
  - 6.3.1 Introduction
  - 6.3.2 Annual Wage Transitions
  - 6.3.3 Weekly Wage Transitions
  - 6.3.4 Property Income Transitions





7.	<u>Full Model Simulation: 1967 to 1971</u>	211
7.1	Introduction	
7.2	Simulation Results: 1967-1971	
7.3	Analysis of the Simulated 1971 Population	
8.	<u>The Policy Block</u>	248
8.1	Evaluating the Effects of Government Programs	
8.2	Policy Algorithms	
8.3	Running the Policy Block	
8.4	Example of Policy Simulation: The Personal Income Tax Algorithm	
9.	<u>Conclusions</u>	259





## APPENDICES

### A. Initial Year State Description

- A.1 Consistency of Probabilities and Process Sequence
- A.2 Detailed Specification of the Individual State Vector
- A.3 Census, SCF Survey and Model Population Tables
- A.4 Survey of Consumer Finance Weighted Work File Magnetic Tape Layout
- A.5 Data Tape Layout for the Initial Year State

### B. The Immigration Block

- B.1 Derivation of Family Size Probabilities
- B.2 Listing of Data
- B.3 Validation Results
- B.4 Computer Program IMMIG

### C. The Demographic Block

- C.1 Parameter Estimation and Calibration
- C.2 Validation of the Demographic Block
- C.3 Initial Population Errors
- C.4 Computer Program DEMOG

### D. The Activity Status Block

- D.1 Calibration of Transition Matrices
- D.2 Validation of the Activity Status Block
- D.3 Computer Program ACTIV
- D.4 Activity Status Block Data

### E. The Market Income Block

- E.1 Definition of the Market Income Block Parameters
- E.2 The Estimation of Transition Matrices
- E.3 Computer Program INCOME
- E.4 Listing of Market Income Block Data





F. The Policy Block

F.1 Computer Program RESULT

F.2 The 1972 Income Tax Algorithm

F.3 The 1973 Income Tax Algorithm





## PREFACE

This report contains full documentation of the POLSIM model constructed as a project of the Effectiveness Division, Planning Branch.

The present report is essentially organized into three tiers. The first comprises chapters one, two, seven, eight and nine. These give the reader an overview of the nature of the model together with some test results and certain conclusions based on the project's experience. The second tier is represented by the full text of the report (excluding appendices). This takes the reader into the detail of the model's structure and the manner in which it was estimated. The third and last tier is represented by the appendices which treat certain structural and estimation questions in greater depth, document the parameter values of the current model, present validation results and list model software.

The model has been run at Statistics Canada on an IBM 370-165 computer. Initial year input tapes for 1967 and 1971 are stored in the Statistics Canada Tape Library together with synthesized or simulated population tapes for the years 1968, 1969, 1970, 1971 and 1972. At the moment of writing this report, the model is run from a source deck but will shortly be available in a more efficient, compiled form.





## 1. INTRODUCTION

### 1.1 Motivation for the Development of POLSIM

The public sector affects both the level and distribution of national income. Until fairly recently concern has largely been focused on questions of the level. This has been reflected in Canada in the development of a number of aggregate economic models which have tended to concentrate on the business sector. Perhaps the best known of these are RDX2 and CANDIDE. Although these models do provide treatment of both the household and government sectors, the manner in which this is done does not lend them to the study of the income distributional consequences of government programs. This is the case because the distribution of income is most meaningfully considered in relation to individual households or families, the recipients of national income, and these models do not maintain sufficient household or family detail. It is clear that any exercise designed to elucidate distributional issues must begin with the household sector portrayed in great particularity, preferably at the level of the individual person.

Of course, given the availability of elaborate sets of microdata, it is possible to develop fairly simple models that enable one to pose "what if" questions for some past period. But this kind of static analysis, although useful, does not allow one to come to grips with a host of questions bearing on income distributional issues which are crucially dependent on time. To be policy





relevant, it is also necessary to have the ability to forecast the workings of a large number of factors, economic and demographic, that bear on the household sector and determine the distribution of income.

It is also necessary to have the capacity to model government programs as comprehensively as possible, and in considerable structural detail. It is not possible to know what any single program will do if we are unable to situate that program in the context of an environment produced, in part, by a number of other programs. Further, it is not enough to model the general direction of the effects produced by a given program as a whole. We are concerned to understand the significance of particular program designs. To do this we require the ability to test the effect caused by an alteration of internal program components.

## 1.2 Microdata Simulation

The term simulation is usually used to describe techniques related to the construction of models or simulators whose operations are intended to resemble the behaviour of actual or potential operating systems. Microdata simulation involves the construction of simulators intended to function in the same manner as operating systems comprised of a large number of basic components or decision units. In carrying out this form of simulation one may employ the technique of endowing microcomponents with behavioural functions and of deriving the consequences (in terms of individual





behaviour) of different environments (specified by a set of independent variables) on the microcomponents. Or, one may employ the technique of stochastic events. In this case microcomponents are not endowed with explicit specifications as to their behaviour under differing environments but rather are thought to behave as particles governed by specified laws of chance. That is, the causal relationships that actually determine behaviour are considered only implicitly, either because they are too complex to be modelled, or because insufficient data exists to estimate the specified model. One is only able to observe that the microcomponents reside in a certain "state" for a period of time and then move to other "states". The precise laws governing individual movement are unknown. What is known is that if a large number of "state" changes or movements are observed, the behaviour in question can be described as if it depended solely on chance.

Events which are assumed to be governed by chance are described by probability distributions. The fact that we are able to specify these distributions means that we know something about the process in question, although we can't fully explain the causal laws underlying it. Changes in the environment take the form of changes in the probability distributions which govern the chance outcomes. In the limit, if we understood the process completely, we could specify these probabilities as either zero or one for particular individuals, depending on the environmental factors. That is, we would have learned enough about the process to eliminate all of the randomness and it would become completely deterministic.



### 1.3 Overview of the Model

The probabilistic analog of a deterministic causal process is the first order Markov-chain. In this kind of process, the future state of the world is completely independent of time periods preceding the present. In the present context, then, micro-component behaviour would be described completely by its present state, and the Markov-chain probabilities that relate the present to the future. In the broadest possible terms this is the POLSIM Model: it is a Markov-chain simulation of individual demographic, labour force, and market income behaviour.

More specifically, POLSIM is an annual microdata model of the Canadian household sector. The basic component of the model is the individual person. Individual persons may be associated into nuclear family units in the model but the prime focus is always maintained on the individual. The model receives, as exogenous input, a specification of the native Canadian population for some year. This specification is made in terms of a number of characteristics (a state vector) for every individual in the initial population.

The individual state vector is comprised of three different sets of characteristics: demographic (e.g., age, sex, etc.), activity (e.g., weeks employed, weeks unemployed, etc.) and income (e.g., annual wages, annual dividend income, etc.). Most of these characteristics change over time. One of the main functions of the model is to effect these changes in the light of





the individual's particular circumstances, as portrayed by his state vector immediately before the change, and conditions prevailing in the socio-economic environment. Two of these changes, death and emigration, in effect destroy an individual state vector.

A second main function of the model is to introduce immigrants into the model "population". This is essentially a problem of completely constructing an individual state vector for each immigrant. This problem does not arise in the case of the native Canadian population, because for this group the state vector is given (see Chapter 2). In the case of immigrants, on the other hand, no such information is available. Last, the model functions to compute the effects of a range of government programs. At the moment these effects are mainly restricted to changes in financial flows, but the possibility exists to extend this treatment to comprehend other effects as well.

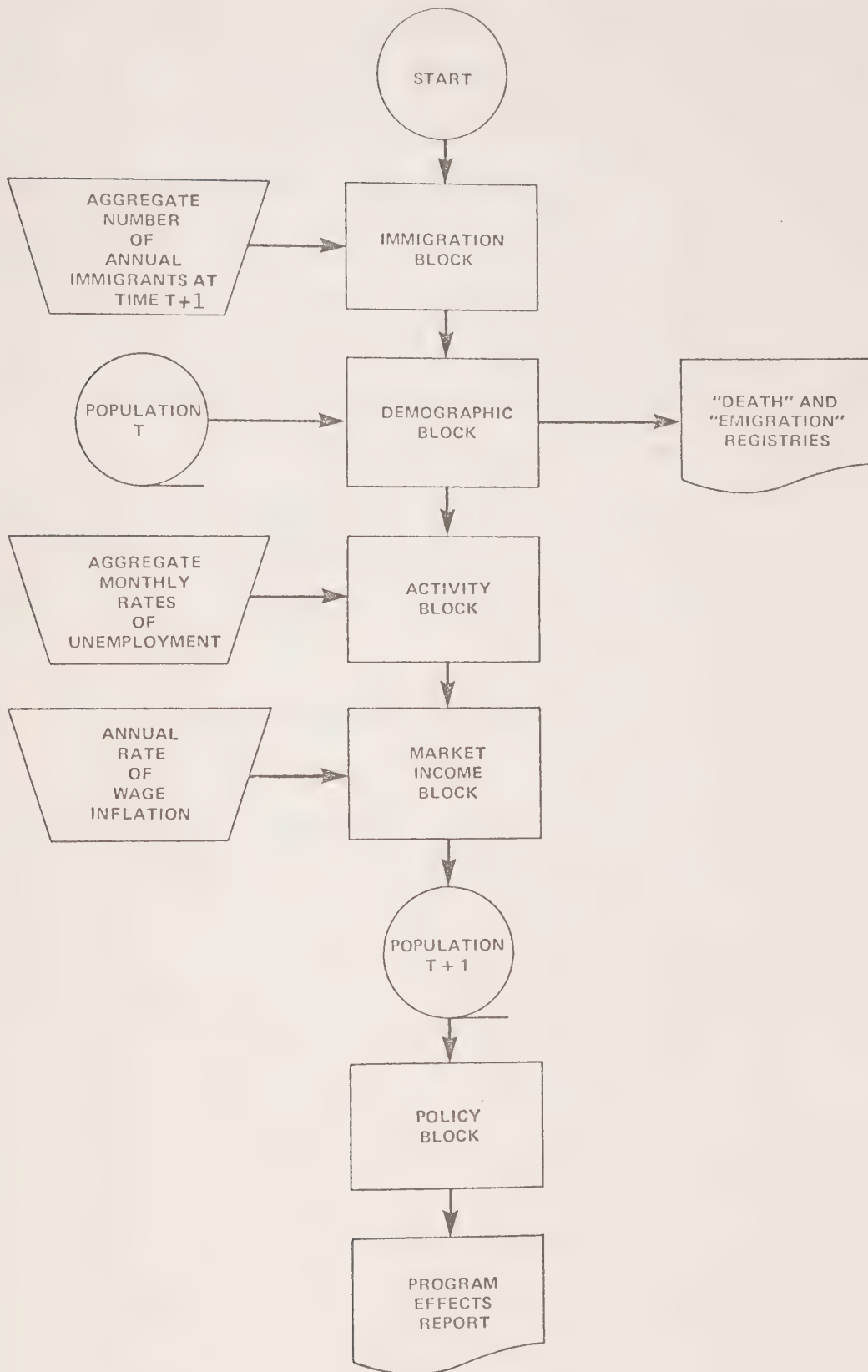
The POLSIM model is constructed as a number of connected blocks: (i) immigration, (ii) demographic, (iii) activity status, (iv) market income, and (v) policy. Each of these blocks contains a series of processes or transformations which operate on individual state vectors to produce annual change. These processes are ordered within blocks. For example, within the demographic block an individual must first pass through the "survival" process before being considered for the "emigration" process, etc. And similarly, the blocks are themselves arranged in the order listed above (illustrated in Figure 1.1). That is, the model requires





Figure 1.1

## THE POLSIM MODEL





that individuals pass through the demographic block before entering the activity block, and so on. This treatment is, of course, artificial. In the real world these processes do not operate in some established sequence but rather simultaneously and continuously. However, no great violence is done to reality provided that we derive consistent measures of the probabilities which "describe" these processes and arrange the process such that those which affect the outcome of other processes precede the latter in sequence. The question of consistency of probabilities and the sequence of the various processes is treated in Appendix A.1.

The purpose of the immigration block is to open the model to international in-bound migrants. This is done, as we have already related, by fabricating the appropriate number and kind of individual state vectors and introducing these into the model population. In a simulation of population and other changes from one year to the next, the immigrants of year  $t+1$  first appear in the model at the beginning of year  $t+1$ . That is, all immigrants are assumed to arrive on January 1. The immigration block receives two inputs exogenous to the model: the aggregate annual number of immigrants to Canada (e.g., 150,000 individuals), and the aggregate rate of unemployment that is assumed to obtain in January of the year being simulated.

The demographic block is concerned with the problem of updating both the native and current immigrant populations through the processes of: (i) emigration, (ii) death, (iii) aging, (iv) divorce/separation, (v) marriage, (vi) fertility, (vii) family dependency, and





(viii) internal migration. The demographic block takes as exogenous input the whole set of individual state vectors which collectively describe the native Canadian population in some given year  $t$ . These are passed through the block one by one, together with the output of the immigration block (i.e., the individual state vectors of all current year immigrants), and updated by the processes mentioned above. The demographic block, then, produces an integrated (i.e., native plus current immigrant) population possessing partially updated individual state vectors (i.e., updated in their demographic characteristics).

The activity block accepts the set of partially updated state vectors from the demographic block and an exogenous input consisting of thirteen monthly Canadian aggregate unemployment rates. Here is the first instance in the model where the socio-economic environment, as distinct from the particular characteristics of the individual (as described by his state vector) influences the nature of the changes to which an individual is subjected. The activity block consists of three processes. The first, a monthly labour force model, places individuals in one of four possible states: school, other non-labour force, employed, and unemployed. Here the time frame of the model shifts, temporarily, from one year to one month. The second process, educational attainment, updates the education status of individuals who have spent the requisite time in school during the simulation year. The last process, type, determines how new labour force entrants (i.e., persons



who either leave school or the non-labour force and join the labour force during the simulation year) will relate to the labour force. This designation segregates, for the purposes of the model, the labour force into two groups - those persons who will be subject to the risk of unemployment and those persons who will not. It also determines whether or not the type of employment the person enters will be such as to allow him a private pension on retirement. The activity block, then, does its job by further updating the set of activity characteristics for all of the individuals in the integrated population. The final update, that of income characteristics, can now be undertaken.

The market income block consists of five processes, one for each of the different kinds of market source income that are identified in the model. These are wages of persons subject to unemployment, income from employment of persons not subject to unemployment, dividend income, other property income, and private pension income. In the case of wages of persons subject to unemployment the market income block generates weekly wage rate change. Annual wage income is then calculated as the product of weeks employed and the updated weekly wage. Persons not subject to unemployment simply have their annual employment income updated directly. In both of these cases the model calculates wage change in real terms and accepts an exogenous specification of wage inflation to convert to money wages. The remaining kinds of income are altered on a pure money change basis. New entrants to the work force are endowed with initial wage rates or annual





wages in a manner which reflects their accumulated human capital. After these initial endowments, further changes in employment or wage income are effected in the same manner as with those persons who are already in the work force. In sum, the market income block accepts partially updated state vectors from the demographic and activity blocks and an exogenously posited wage inflation rate; it then completes the update of individual state vectors.

The model population is now fully updated for one year. All changes that will occur between year  $t$  and year  $t+1$  have been made. It remains to pass the population through the policy block in order to calculate the effects of defined government programs for the year  $t+1$ . The policy block of the model consists of a number of algorithms designed to simulate certain of the effects of these programs, (e.g., taxes paid, welfare benefits received, etc.). These algorithms may be thought of as a series of processes which accept the state vectors of individuals comprising a single family, together with a specification of a particular government program, as inputs and which then calculate program effects as outputs.

#### 1.4 Running the Model

One pass through the entire model produces an update of the population for a single year. A succession of passes, with a given set of time labelled exogenous inputs, produces a unique time track of the population. Any change in any one of the several exogenous inputs



will result, of course, in another time track. However, in terms of actual computation, it is only necessary to create a completely new time track from the beginning of the block in which the exogenous input has been altered. This results in a considerable saving of both effort and computing cost.

The demographic block may be run separately to describe a number of "demographic" tracks depending on the assumptions made with respect to aggregate immigration. These "demographic" tracks may in turn be run through the activity block to generate a number of "demographic-activity" tracks, depending on the assumptions made with respect to aggregate unemployment. Finally, one or more of these "demographic-activity" tracks can be expanded into a larger number (dependent on the number of assumptions made with respect to wage inflation) of "demographic-activity-income" tracks.

#### 1.5 Examples of Simulation Experiments Feasible Using POLSIM

In section 1.1 above we mentioned that one of the main motivations for the development of POLSIM was the desire to evaluate the distributional consequences of government programs (mainly tax and transfer programs). At the moment of writing this report, a fairly large number of government programs have been modelled (see Chapter 8 below). The redistributive consequences of any complex of these programs may be examined in the context of one or more time tracks.





For example, the cost and initial incidence of a given negative income tax could be examined over the period, say, 1973-1976 under different assumptions in respect of immigration, unemployment and wage inflation. Or, the yield and initial incidence of the indexed personal income tax may be estimated for the same period and under the same set of assumptions. Alternatively, we may be interested in the interactions between two or more programs. Say, for instance, the number of persons subject to cumulative marginal tax rates of 50% or more from all sources. Another simulation experiment could be the determination of the persistence of individual poverty over time or the relative efficacy of direct transfers as opposed to other measures for poverty alleviation.



## 2. INITIAL YEAR STATE DESCRIPTION

In section 1.3 above we indicated that POLSIM requires the specification of a model "population" for some given year. In the present chapter we shall describe this initial population more fully. We will begin by looking more closely at the composition of the individual state vector to see the particular characteristics which it contains. We shall then briefly consider a number of alternative data sources for initial population specification, and end the chapter with a discussion of the initial model population.

### 2.1 The Individual State Vector

In section 1.3 we described in very broad terms the nature of the individual state vector, that is, the fact that it contains demographic, activity and income characteristics. The complete vector in its most detailed form contains 23 characteristics as shown below:

1. Previous Year Family Unit Number (LYUNIT)
2. Family Unit Identifier (UNIT)
3. Province (PROVIN)
4. Size of Family (SIZE)
5. Census Family Relationship (DEPNKY)
6. Marital Status (MSTAT)
7. Age (AGE)
8. Sex (SEX)
9. Major Source of Income (MAJSIN)
10. Weeks in School (WKSCHL)
11. Weeks employed (WKEMP)
12. Weeks unemployed (WKUNEM)
13. Weeks in the non-labour force (WKNLF)
14. Education (EDUCTN)
15. April Activity Status (YRACT)
16. Weight (WEIGHT)
17. Employment Category (TYPE)
18. Employment Income (EMPINC)
19. Interest and Other Investment Income (INTRST)





- 20. Dividends (DIVDNS)
- 21. Retirement Pension, Superannuation, and  
Annuities (RETIRE)
- 22. Other Money Income (OTHER)
- 23. Total Income (TOTAL)

The names of most of the characteristics listed above explain the meaning of the particular characteristic. Some, however, are less evident. In particular, "Family Unit Identifier" is simply a unique number which all members of a given nuclear family have in common. "Weight" is the sample weight which an individual has. This is explained more fully in section 2.3 in the discussion of the initial model population. The entire individual state vector is described in greater detail in Appendix A.2.

## 2.2. Alternative Sources of Initial Year Data

A number of sets of household sector microdata do exist which could lend themselves to the specification of the model population for some initial year. We will comment very briefly on these from the point of view of population coverage, the nature and quality of data and the frequency of issue.

### 2.2.1 Department of National Revenue (Taxation) Tax Analysis Data Base

The Tax Analysis Data Base is an annual stratified sample consisting of 1.25% of filed T1 Short and T1 General tax returns. The Tax Analysis Data Base is, naturally, heavily disposed toward information concerning income, deductions, exemptions and other income taxation data, but certain other data are also carried.



The annual nature of this DNR data set means, of course, that it is very timely. Microdata projections can be readily checked. And, the possibility of continuous updating of the initial year is extremely useful for the accuracy of microdata projections. Furthermore, the quality of income data for upper income level persons is probably superior to any other available. However, two serious limitations attach to this data set. First, the population coverage is obviously quite incomplete. Only those persons who file income tax returns are represented. This means that many low income persons are not covered. Second, sample records are those of individual persons as distinct from families. Although it is possible to infer something about the family of the tax-filer, it is not possible to identify two or more tax-filers who belong to the same family unit.

#### 2.2.2 Unemployment Insurance Commission Data Base

The Unemployment Insurance Commission Data Base is a 2% sample of all persons with social insurance numbers. It consists of data describing the demographic, financial, and employment characteristics of approximately 2% of the Canadian working population. The sample comprises approximately 250,000 individual records, and was compiled from two main sources: Statistics Canada and the Department of National Revenue. The Statistics Canada files contained data derived from UIC administrative records, as well as information on occupation and industry. The DNR records supplied income information. Data sets for persons who are not unemployment insurees



or who do not file income tax returns are not complete. The data base contains information for the years 1965-1971.

The UIC data base suffers from the same weaknesses as does the DNR data base. Population coverage is restricted to persons with social insurance numbers, and family records cannot be inferred from the given individual records. In addition, there is little prospect that this data base will be updated on a regular basis, since it was originally developed for a special purpose task. It is thus not a promising source as a base population for a micro-simulation model.

### 2.2.3 The Census

The census is of course the most comprehensive data base in existence in Canada. Unfortunately, for purposes of micro-simulation, it has several defects. First, it is too comprehensive. It is just not practical to simulate an entire population. Instead, what is wanted is some sample of the population. A sample of the census records could be taken, of course, but this would involve considerable expense, problems of weighting, and extensive computer programming. Second, the census is untimely. Since it is only taken every 10 years, it rapidly loses its usefulness as time from the previous census elapses. And finally, it is unwieldy. Since any simulation model can at best focus on only a limited number of variables, extensive software would have to be developed simply to cut the individual census records down to manageable size. For all of these reasons, the census is not an ideal source as an initial year model population.





#### 2.2.4 Survey of Consumer Finance

The data carried in the Survey of Consumer Finance is extensively documented in Appendix A.5. In the recent past the Survey has been conducted every two years. In future, it will be taken annually, but every second survey will be small scale and of a specialized nature so as to make it unsuitable for initial year model population specification. For the present purpose, therefore, we may only regard each second survey as useful.

The SCF has the advantage of more extensive population coverage than either of the DNR or UIC microdata sets mentioned above. Coverage extends to all of the population of Canada with the exception of (i) persons resident in the territories, (ii) persons resident on Indian reservations, and (iii) persons resident in institutions (e.g., prisons, mental hospitals, etc.). The Survey carries a wide variety of individual income and other information (see the complete documentation of SCF data in Appendix A.5) capable of organization on either an economic or census family basis. The quality of this data is generally very high (see Appendix A.3 and Appendix A.4 for comparisons with other data), and its availability every two years makes it timely. All things considered it seemed best to utilize the Survey of Consumer Finance as the source for the initial model population data.



### 2.3 The Initial Model Population

In most cases the data for an individual's initial year state vector are simply taken directly from the Survey of Consumer Finance (see Appendix A.5 below). In some cases, however, it is necessary to derive this information from other sources. The first instance where this occurs is the case of persons under 14 years of age. The Survey of Consumer Finance does not record sex, education or activity status of these persons. Sex status must be assigned by simulation using the Monte Carlo technique and known probabilities. Education status and activity status, on the other hand, are inferred in simple fashion from age. That is, it is assumed that the individual enters primary school at age 6 and proceeds mechanically through the various grades so that age 6 is equivalent to "grade 1" and age 13 is equivalent to "grade 8".

A similar problem with activity status exists for persons 14 years and over. If the individual is not in school according to the SCF data, his activity status (employment, unemployment, or non-labour force) can be inferred directly from his labour force status as recorded in the SCF data. If, on the other hand, the person is shown to be in school, his activity status is determined on the basis of age, province, and education level. People in high school are assigned grades according to age, the sequence related above being continued, (i.e., age 14 corresponds to "grade 9" and so on). In Ontario, allowance is made for the attainment of grade 13; other provinces are assumed to have secondary





education up to and including grade 12. Persons in university are placed in either 6th year university or 2nd year university, depending on whether or not they are shown by the Survey of Consumer Finance to possess a university degree.

Each of the 23 characteristics, discussed above, taken together "describe" one individual. The entire Canadian population could in principle be "described" in the context of the model by a large number of individual state vectors - one for each individual in the population. This would, however, be a very inefficient procedure. A much more economical means of accounting for the entire population is by way of a representative sample. In this case, each individual state vector in the model stands for or represents a larger number of "identical" persons in the real population.

The model requires, as initial input, a weighted sample of individual state vectors which describe the population of Canada for some year. At the time of writing this report we have been working with two such initial years, 1967 and 1971. The Survey of Consumer Finance undertakes a geographically stratified sampling of Canadian households in April of the year following the survey year (e.g., April 1968 for 1967 or April 1972 for 1971). The sample included 37,985 individuals over fourteen years of age and in receipt of cash income in 1967 and 43,039 individuals aged fourteen years or more who were in receipt of cash income in 1971. Over time, of course, as more current data becomes available, initial year state descriptions for more recent periods can be constructed. It is important to be able to continually update the initial year as a guard against the generation of inaccurate projections.



Individual state vectors are weighted by the Survey of Consumer Finance in order to calculate population estimates. These weights are unequal, running from 30 up to 3,000 in increments of 10. In order to conform to the logic of POLSIM, individual state vectors must possess equal weights. It was necessary, then, to adjust the SCF samples in such a way as to produce equally weighted records. We adopted a common weight of 50 (i.e. one individual in the sample represents 50 in the real population). All SCF records first had their weights randomly rounded to be multiples of 50. Next, with all record weights some multiple of 50, we replicated each record a number of times equal to the quotient of the weight divided by 50. For example, if the original SCF record had weight 520, it was randomly rounded to either 500 or 550 which in turn yielded  $\frac{500}{50} = 10$  or  $\frac{550}{50} = 11$  identical records, each of weight 50.

More explicitly, let the original SCF weight be  $W$ . We wish to randomly round this weight to a multiple of a positive integer  $m$ . We have then,  $W = mg + r$  where  $r = 0, 1, 2, \dots, (m-1)$  and  $g$  is some positive integer.  $W$  is rounded by Monte Carlo simulation to be either  $\bar{w} = mg$  with probability  $f$ , or  $\bar{w} = mg + m$  with probability  $(1-f)$ . In order to be unbiased we require the expected value of  $\bar{w}$  to equal  $w$ . That is we require,

$$E(\bar{w}) = f(mg) + (1-f)(mg + m) = W$$

$$\text{or } E(\bar{w}) = mg + (1-f)m = W$$

which yields  $(1-f) = \frac{r}{m} = \text{Prob } (W \text{ will be rounded to } mg + m)$  and  $f = 1 - \frac{r}{m} = \text{Prob } (W \text{ will be rounded to } mg)$ .



For our case,  $m = 50$ , any integer  $W$  divided by 50 will give a residual  $r = 0, 1, 2, \dots, 49$  and the ratio  $\frac{r}{m}$  is the probability that this residual will be increased to 50.

The process of replicating sample state vectors increases the number of sample individuals from 79,479 to 397,960 in 1967 and from 79,528 to 425,864 in 1971. This larger sample is useful for Monte Carlo simulation, of course, since simulation errors are thereby reduced. A potential difficulty, however, also attends the replication process. This stems from rounding errors in the process described above. The extent of this error can be assessed by comparing population estimates produced from the SCF unequally weighted and the POLSIM equally weighted samples. A comparison of Tables 2.1 and 2.2. reveals that the rounding error is not great. (See also the more detailed tables of Appendix A.3).

There remains the question of the adequacy of the base year data as a description of the Canadian population for the relevant year. An excellent simulation model will obviously produce bad projections if the initial year state description is poor. A comparison of Table 2.2 with Table 2.3 shown below provides a general idea of the adequacy of the initial population coverage. (See also the more detailed comparisons of Tables A3.5 and A3.10, Appendix A.3.) In general, the 1971 SCF, and as a consequence our 1971 initial year model population, understates the 1971 Canadian population by one percentage point. This understatement is worst in the case of older age groups (i.e. persons over eighty years of age). The Province of Saskatchewan population is, interestingly enough, also considerably understated.





TABLE 2.1

Survey of Consumer Finance Population  
of Canada, April 1, 1972

Age Group	Male	Female	Total
0-9	----- 3,992,460 -----	-----	3,992,460
10-14	----- 2,318,960 -----	-----	2,318,960
15-24	1,952,000	1,939,190	3,891,190
25-44	2,724,410	2,758,460	5,482,870
45-64	1,944,560	1,953,870	3,898,430
65-95	800,620	914,160	<u>1,714,780</u>
			21,298,690

Source: Table A3.1, Appendix A3.

TABLE 2.2

1971 Initial Year Model Population of Canada

Age Group	Male	Female	Total
0-9	2,042,750	1,947,850	3,990,600
10-14	1,179,850	1,138,850	2,318,700
15-24	1,952,800	1,940,000	3,892,800
25-44	2,722,650	2,757,200	5,479,850
45-64	1,945,550	1,952,600	3,898,150
65-95	799,200	913,900	<u>1,713,100</u>
	<u>10,642,800</u>	<u>10,650,400</u>	21,293,200



TABLE 2.3

Census Population of Canada  
June 1, 1971

---

Age Group	Male	Female	Total
0-9	2,082,025	1,988,135	4,070,160
10-14	1,181,450	1,129,285	2,310,735
15-24	2,016,210	1,987,545	4,003,755
25-44	2,747,395	2,668,545	5,415,940
45-64	1,986,425	2,036,905	4,023,330
65-95-	<u>781,865</u>	<u>962,535</u>	<u>1,744,400</u>
Total	10,795,370	10,772,950	21,568,320

Source: Table A3.3, Appendix A3.

The April labour force status of that portion of the population captured in the Survey of Consumer Finances is reliably reported since the SCF is tied to the April Labour Force Survey. We can have, therefore, considerable confidence in these data though not so much in the number of weeks of employment, unemployment, etc. reported for the preceding year. Fortunately the latter data is not utilized by POLSIM.



Assessment of the adequacy of the Survey of Consumer Finances for purposes of measuring income can only be carried out at the most aggregate level. Table 2.4 compares SCF income data with adjusted national accounts personal income data. So far as one can tell from this comparison total wages and salaries are fairly accurately captured. The other categories of income are more or less badly underreported. It is of interest that the SCF performance is quite variable over time in this last respect.

TABLE 2.4

Comparison of Survey of Consumer Finance Income\*  
Estimates of National Accounts Adjusted Personal Income

(SCF as percent of Adj.NA)

<u>Item</u>	<u>1967</u>	<u>1971</u>
Wages and Salaries	101.8	102.9
Non-farm Income from Self-Employment	70.7	59.6
Farm Income	95.1	82.4
Interest, Dividends and Miscellaneous Investment Income	49.4	70.0

\* Individual series.

Source: Unpublished data, National Income and Expenditure Division and Household Statistics Branch, Statistics Canada.





### 3. THE IMMIGRATION BLOCK

#### 3.1 Purpose and Overview

The objective of the immigration block is to increment the base population in a given year by the total number of immigrants that will arrive in that year. This entails two distinct problems: the projection of the total number of immigrants who are to arrive in the given year, and the synthesis of individual state vectors for each of these individuals. The present version of the model does not attempt to determine projections of the total number of immigrants. Such projections are simply assumed by the model to be exogenously determined. The immigration block is therefore concerned principally with the construction of individual state vectors for a pre-determined number of people.

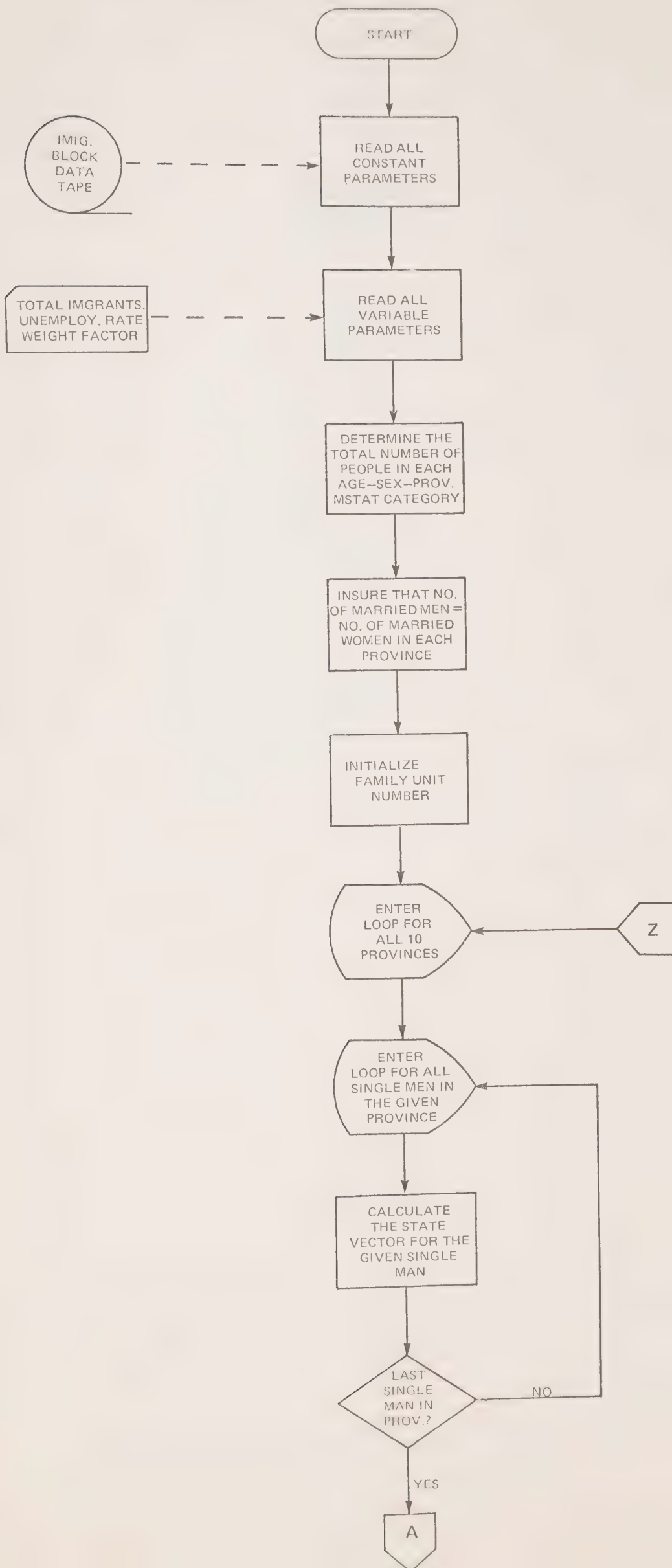
Since the arrival of immigrants is distributed over all 12 months of the year, there is some question as to when exactly they should be added to the population base. It is assumed, for purposes of the model, that immigrants who arrive throughout a given year will become part of the population base for the whole year. This in effect means that all immigrants in a given year arrive on January 1 of that year, and thus ignores the difficulty of handling people who are only present in the population for part of the year.

#### 3.2 General Structure

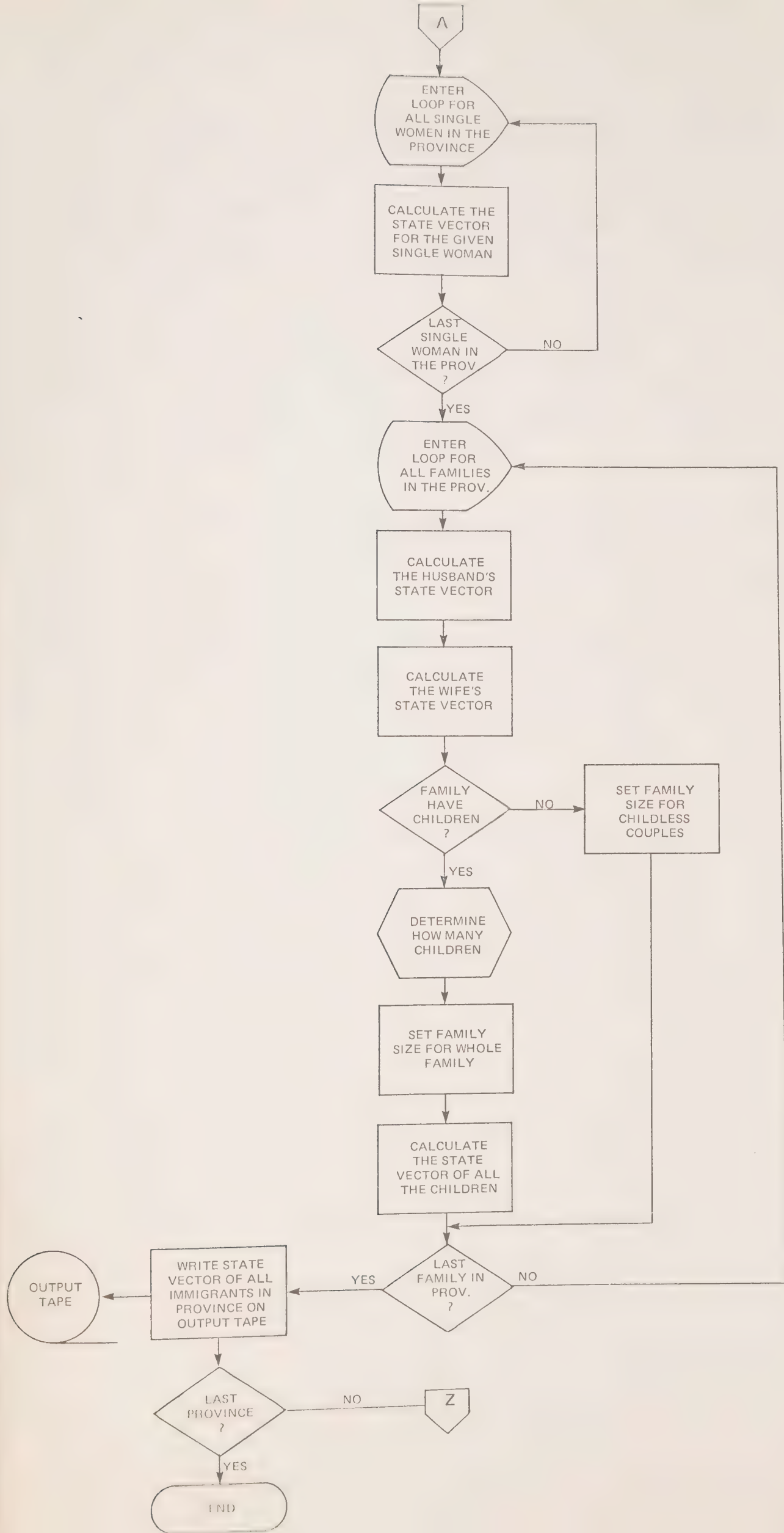
The general structure of the immigration block may be examined with reference to the flow chart in figure 3.1. The model begins by reading in all of the constant



FIGURE 3.1 - THE IMMIGRATION BLOCK











parameters. These include distributions of the immigrant population over age, sex, marital status and province, income distributions of new immigrants, and so on.

The following three exogenous input parameters are then read in:

1. W1 - The weighting factor that applies to the base population. This is an integer; and in the present version of the model is 50.
2. RATE - This is the national mean unemployment rate in decimals for the year in question.
3. TOTIMG - the total number of immigrants that are expected to arrive in the year simulated (including children). The number will be an integer in the range of 100,000 to 150,000.

Once all of the input parameters have been read into the model, the determination of the immigrant population itself can begin. The first step is to calculate the total number of people in each of 240 age-sex-marital status-province classes. These classes, it should be noted, consist only of adults. The numbers of children in given classes are determined after married couples are formed. The number of adults in a given class is the product of the total number of immigrants arriving and the probability of being in a given class, divided by the weighting factor.



The above procedure does not guarantee that the number of married men in a given province will equal the number of married women in that province. The reason for this arises from the fact that no data is kept by which married people can be linked to their spouses, and because the data that does exist is such that the number of married women immigrants always exceeds the number of married men immigrants. Two means of dealing with this problem are possible. The first is to assume that the excess females consists of either (a) widows who list their marital status as married, or (b) women who are planning to join husbands who arrived in earlier years, or (c) women whose husbands will arrive in subsequent years. It would then be necessary to identify in which group each of the excess females belongs, and to try to come up with some reasonable method of "disposing" of them from there. The second approach would be to adjust the raw totals so as to equate married males with married females. This latter method was the one adopted, because the problem affects only a small percentage of the immigrant population, and because inadequate knowledge precludes any reasonable handling by the more sophisticated approach.

Since the data indicates that the largest group of married women falls into the second age group (20-30) the adjustment procedure was simply to delete the excess from this group. The number of excess married women was approximately 3000 (2.5% of the total immigrant population), so we are assured that this procedure does not constitute a gross distortion.



The family identifier number is then initialized at 1. These identifier numbers will be changed so as to fall into the initial population sequence when the immigrant population is merged with the base population. The procedure ensures that all identifier numbers will be higher than those for the base population.

The province loop is now entered. State vectors are determined for all the people in a given province, and these are then written out on tape. State vectors for single males are calculated first. The procedure is to consider one individual at a time, calculating his entire state vector (cf. section 3.3 below). The next individual is then considered, and so on, until all single males have been dealt with. An analogous procedure is then repeated for all single females, and finally for all members of families. The family loop includes the calculation of the number of children a given family will have, as well as these children's respective state vectors.

Once all the immigrants in a given province have been created, their state vectors are written out, and the calculations are repeated for the next province. The final output from the immigration block consists of state vectors describing all of the immigrants in each of the ten provinces. These state vectors are now ready to enter the demographic block along with the remainder (i.e. non-immigrant portion) of the population.





### 3.3 Calculation of the State Vectors

All of the 23 elements in the individual state vector (cf. section 2.1) must be assigned to each of the created immigrants. In many cases, assignment of a particular element is straightforward. (Sex and marital status for single men, for example.) In other cases, however, the elements must be determined in some probabilistic fashion, or else inferred from some other element or some other set of data. These latter cases will be discussed below.

Age presents the first difficulty. We know that each individual must fall into one of six age groups, and we know further how many individuals must be in each group (from the age-sex-province-marital status breakdown arrived at above). The procedure is to simply keep track of how many individuals have been added to a given age group, beginning with the first, until that group is filled up; and then to proceed to the next group, and so on. The age assigned will be the midpoint of the relevant group (i.e. 17, 25, 35, 45, 57, or 70).

It is assumed that the number of weeks employed for all immigrants is 52, and that the number of weeks in each of the other activity states is zero. The only exception to this rule is for married women in the non-labor force. They are assigned zero weeks of employment and 52 weeks in the non-labor force. These assumptions are made only for the purpose of ensuring that weekly wage rates can be calculated if necessary in the Activity



Block. The actual number of weeks that the immigrant will in fact spend in each of the four states will be determined by the Activity Block, given his January activity status in the year of immigration. The January activity status is determined for all single people (excluding children) and married men depending on the mean unemployment rate which is entered as exogenous input and the assumption that they must either be employed or unemployed. Married women, on the other hand, are assumed to be capable of entering the non-labour force category as well. Their activity status is determined depending both on the participation rate of married females and the January unemployment rate. The "Employment Category" for labour force participants thus determined is assumed to be Class B (TYPE-25) unless the person earns more than \$9,000, in which case he is assumed to be Class A (TYPE-15).<sup>\*</sup> The education category is determined by sampling from an education-age-sex distribution, and initial employment income is determined from the income-sex-marital status-age distribution (cf. data description, section 3.4). All other sources of income are assumed to be initially zero. The major source of income is therefore wages and salaries.

In the family section it is necessary to create family units: husbands, wives, and children. A preliminary step toward creating these family units has already

---

\* TYPE is the variable which defines the way in which a person relates to the labor force. TYPE 15 persons are those who are assumed to never become unemployed and who will receive a pension on retirement. TYPE 25 persons will also receive a pension on retirement but differ from TYPE 15 individuals in that they are assumed to be subject to unemployment. These definitions are explained more fully in Appendix A.2.



been taken. As discussed above, the number of married females in a given province has been adjusted so as to equate with the number of married males. The difficulty now is to pair these couples up. This is done solely on the basis of age: the first woman in the first age group is "married" to the first man in the first age group, the second to the second, and so on, until the last woman in the last female age group is paired with the last man in the last male age group. Because the number of males in a given age group does not necessarily equal the number of females in that age group, there will be some men married to women in lower (or higher) age groups, depending on the relative numbers in each group. But because the totals over all age groups in a given province are equal for both sexes, all people will eventually find a spouse.

The method by which children are assigned to parents proceeds as follows. Children are assumed to fall into three age groups: 0-9, 10-14, and 15-19. The mothers who are then allowed to have these children are themselves restricted to three age groups: 20-30, 31-40, and 41-50. That is, all women over 50 and younger than 20 are assumed to have no dependent children. (It should be noted that this probably produces some distortion. But in the absence of data linking family records, any assumption will necessarily be somewhat arbitrary.) It is then assumed that mothers in a given age group will only have children from the same ordinal group. That is, mothers from the first mother age group will only have children from the first child age group, and so on. The number of children in the given age group that a woman will have is then determined by





sampling from a probability distribution. This number may be 0, 1, 2, or 3. If more than one child is assigned it is assumed that they are one year apart in age. The maximum age is assigned to the first child, and lower ages to subsequent children. These maximum ages are respectively 7, 13, and 18, depending on the age group being considered. The sex of the child is determined randomly, assuming half will be boys and half girls. The income of children is assumed to be zero. Family size for all the family members is then set after all the children in the family have been created.

### 3.4 Parameter Estimation

#### 3.4.1 Demographic Variables

The first set of parameters in the immigration block is the distribution of new adult immigrants over age, sex, marital status, and province of residence. More specifically,  $PEOPL(I, J, K, L)$  is the number of adult immigrants of age  $I$ , sex  $J$ , and marital status  $K$  arriving in province  $L$ . The codes for each index are as follows:

AGE =  $I$  = 1 for 15 - 19  
                  2 for 20 - 30  
                  3 for 31 - 40  
                  4 for 41 - 50  
                  5 for 51 - 65  
                  6 for 65 +

SEX =  $J$  = 1 male  
                  2 female

Marital Status =  $K$  = 1 Single  
                                  2 Married



Province = L = 1 Nfld.  
 2 PEI  
 3 NS  
 4 NB  
 5 PQ  
 6 ONT  
 7 MAN  
 8 SASK  
 9 ALTA  
 10 BC

This raw data is derived from the landing records of 1971 immigrants\* and was obtained from the Information Analysis unit of the Programs and Procedures Branch of the Department of Manpower and Immigration. Given this data, and the total number of immigrants arriving in 1971 (121,900), it is possible to derive the probability that an adult immigrant will be in a given age-sex-marital status-province class (see Appendix B for this distribution). At present, it is assumed these probabilities are stationary over time.

A clear deficiency in the calculation of this distribution is that it is based on data from a single year, 1971. A superior method would be to estimate the distribution, from time-series data, as functions of economic conditions both in Canada and abroad and possibly other variables as well. This would be a major study in itself, feasible in a future version of the POLSIM model.

#### 3.4.2 The Assignment of Children

The assignment of children to families is hampered by the fact that the raw data available does not link

---

\* Landing records are collected from all immigrants upon their arrival in Canada. These contain information on the person's age, sex, marital status, etc. The Department of Manpower and Immigration collects all of these records, and stores them on magnetic tape.



children to their parents. This limitation is further exacerbated by questions of pure theory: in principle, one desires a distribution of different family sizes corresponding to differing ages of family members (for example, how many children and of what ages is a 35 year old mother likely to have?). It is clear that the number of possible combinations here is very large; so large, in fact, that a distribution calculated on the basis of the immigrant population would probably lack statistical significance, even if it could be derived.

The data that is available (from Immigration Landing Records) consists of a breakdown of the number of children arriving in 1971 by age, sex, and province. The problem is to assign these children to parents in some reasonable way. The method by which this is done is described below.

As discussed previously, it is first necessary to make the following assumptions:

- (a) wives older than 50 or younger than 20 have no dependent children;
- (b) all children in the 15-19 age group are assigned to mothers in the 41-50 age group;
- (c) all children in the 10-14 age group are assigned to mothers in the 31-40 age group;
- (d) all children in the 0-9 age group are assigned to mothers in the 20-30 age group;





- (e) all women have no more than three children, and there exists some probability of their having a given number (0, 1, 2, or 3).

The problem then reduces to calculating the probability that a woman in a given age group and a given province will have a given number of children. That is, it is desired to calculate CHILD (I, J, K), the cumulative probability that a woman in province I and age group J will have K children. The indices are as follows:

I =	1	Nfld.
	2	PEI
	3	NS
	4	NB
	5	PQ
	6	ONT
	7	MAN
	8	SASK
	9	ALTA
	10	BC
J =	1	20-30
	2	31-40
	3	41-50
K =	1	0 children
	2	1 child
	3	2 children
	4	3 children

In general, the way in which these probabilities were derived is as follows (the details and actual distributions are presented in Appendix B).

Let Y = number of wives in a given age group  
in a given province.

X = number of children in the province to  
be assigned to women in this age  
group.

$p_i$  = probability of a wife in this age  
group and this province having i  
children.



$$\text{Then } p_0 + p_1 + p_2 + p_3 = 1 \quad (1)$$

$$\text{and } E(\text{children}) = p_1 Y + 2p_2 Y + 3p_3 Y = X$$

where  $E$  is the mathematical expectation operator

$$\text{or } p_1 + 2p_2 + 3p_3 = \frac{X}{Y} \quad (2)$$

We thus have two equations in 4 unknowns. This problem is resolved by a priori determination of two of the probabilities by inspection of the data (see Appendix B). The remaining two probabilities are then determined by simultaneous solution of equations (1) and (2). In this way the entire array, CHILD (10, 3, 4), is determined.

#### 3.4.3 Participation Rate for Married Women

The labor force participation rate for married women is derived from the data on the "Intended Occupations of Male and Female Immigrants Admitted to Canada 1971". This data shows that the number of working married females was 5407 in 1971, out of a population of 26,740 married women. The participation rate is thus  $5407 \div 26,740 = .2022$ . (Cf. 1971 Immigration Statistics).

#### 3.4.4 Employment Income Distributions

The cumulative income distribution of certain classes of new immigrants is derived from the longitudinal survey of new immigrants that is currently being conducted by the Department of Manpower and Immigration. This survey follows three cohorts of immigrants through their first three years in Canada. These cohorts consist of 5,962 people from the 1969 immigrant population (a 3.7% sample), 5,338 from the 1970 population (a 3.6% sample), and 5,368 from the 1971 population (a 4.4% sample). Data from each cohort is collected in



four questionnaires. The first is sent out 6 months after the immigrant's arrival, the second at 1 year, the third at 2 years, and the fourth at 3 years. The questionnaires seek to determine the demographic characteristics of the immigrant, his employment and income experience, his social adaptation, and his residential mobility. Once the questionnaires have been returned, they are linked to the immigrant's Landing Record and his Immigrant Assessment Record (if the latter exists). All of this data is then available for compiling particular distributions.

The survey has not yet been completed, and consequently the data is at best tentative. Better distributions will become available some time in 1975, when the survey will be completely finished and tested.

The income distribution compiled for the present version of the model is  $DOL(I, J, K, L)$ . It is the cumulative probability that a person of sex  $I$  ( $1=M, 2=F$ ), marital status  $J$  ( $1=$ single,  $2=$ married), and age  $K$  will be in income group  $L$ .

The age groups are as follows:

1	for	15-19
2	for	20-30
3	for	31-40
4	for	41-50
5	for	51-65
6	for	65+



The average income in each group is:

1	\$ 750
2	\$ 1500
3	\$ 2500
4	\$ 3500
5	\$ 4500
6	\$ 5500
7	\$ 7000
8	\$ 9000
9	\$12500
10	\$18000

The actual distribution is given in Appendix B.

#### 3.4.5 Education Distributions

The education distribution of new immigrants is stratified by age and sex. BOOK (I, J, K) is the cumulative probability that a person of sex I (1=M, 2=F) in age group J will be in education class K. The age groups are the same as in all the other arrays described previously.

The education classes are:

K = 1	Completed Elementary
2	Some High School
3	Completed High School
4	Some College or University
5	University Degree

Like the income data, these distributions are derived from the longitudinal survey of new immigrants. The actual numbers are given in Appendix B.





### 3.5 Validation of the Immigration Block

The validation of the immigration block consisted of an attempt to simulate the 1971 immigrant population. The total number of 1971 immigrants (121,900) was entered as exogenous input and then the simulated population was produced. This total of 121,900 corresponds to a total of 2,438 simulated individuals at a weighting factor of 50. The number of histories actually created was 2,404, the difference being attributable to the fact that the number of children created in any given family is determined probabilistically and the fact that the number of married women was slightly reduced so as to equate with the number of married men.

Once all of the individual state vectors were calculated various distributions were derived from these simulated vectors and then compared with the corresponding actual distributions. A summary of these comparisons is presented in table 3.1, in which the immigrant population, both simulated and actual, is broken down by marital status and province. As the table indicates, the simulation performed quite well. The more detailed tables in Appendix B, which compare other marginal distributions, also demonstrate that the simulation is on the whole quite successful. It should be noted, however, that this does not imply that the simulated joint distribution over all 23 state variables will be in close agreement with the actual joint distribution over these variables. The reason for this is that even if all the marginal distributions were independent of one another, there are simply not enough individuals created to adequately match all of the cells that 23 joint variables produce. (Note that if



TABLE 3.1

Summary Validation of the Immigration Block

Province	Married		Single	
	Simulated	Actual	Simulated	Actual
NFLD.	500	410	200	189
P.E.I.	0	82	0	41
N.S.	800	814	482	550
N.B.	600	510	200	247
P.Q.	9500	7394	7400	7394
ONT.	24100	23988	22000	21968
MAN.	2000	1980	1600	1707
SASK.	600	584	400	375
ALTA.	3300	3454	2350	2432
B.C.	7900	8000	5300	5324

Source: Table B.4.1 and B.4.2

each state variable was dichotomous, which is the minimum possibility,  $2^{23}$  cells would exist). But this is not really a difficulty for the present model. Since all of the marginal distributions are made conditional on as many relevant variables as existing data permits, it can be held with some confidence that the simulated population adequately represents the actual immigrant population.



#### 4. THE DEMOGRAPHIC BLOCK

##### 4.1 The Demographic Block Model

###### 4.1.1 Introduction

The purpose of the Demographic Block is to update the demographic variables of the individual state vector. These variables are the individual's province of residence, his family size, his dependency status (whether he is a family head, a wife, or a dependent child), his marital status, and his age. The Demographic Block also determines whether or not a family will emigrate. Since the basic time unit of the POLSIM model is one year, the updating of these state variables is on an annual basis. The demographic block receives the set of individual state vectors which describe the native Canadian population for some year  $t$  plus immigrants for the year  $t+1$ . The problem is then to determine the necessary changes in the demographic characteristics of the population over the time period  $t$  to  $t+1$ .

The Demographic Block is the first of three blocks in the POLSIM model which alter individual state vectors. The other two, the Activity Block and the Market Income Block, update the activity variables (weeks employed, education, etc.), and the market income variables, respectively. The Demographic Block differs from these latter two in three ways. First, it eliminates and creates individual records or state vectors. New records are created in the Demographic Block through





the simulation of births, while others are eliminated through the processes of death and emigration. The second difference is that the Demographic Block proceeds in two distinct phases. The other two blocks mentioned are somewhat simpler in this respect. They require only a single sequential processing of all individual records. Two phases are necessary in the Demographic Block because of the need to simulate marriage. Since individuals in the base year population must marry other individuals in the same population, it is necessary to first determine all of the "marriageable" individuals. Only then can these individuals be paired. The final difference is that the Demographic Block deals with family records, whereas the other two blocks work with individuals only. The reason for this is that demographic variables are inherently family variables; that is, dependency status, family size, etc., can clearly only be determined within the context of a whole family.

The logical structure of the Demographic Block, and the details of the various processes it entails, are presented in the two following sections.

#### 4.1.2 Structure of the Demographic Block

The Demographic Block consists of two distinct phases. The input to the first phase consists of the file containing the native Canadian population of the previous year together with the population of new immigrants for the current year. The first phase of



the block then produces three output files. The first output file contains the records of all individuals who either emigrate or die during the year being simulated (i.e., the emigration and death registries). The second output file contains the records of all individuals who have been determined to be eligible for marriage during the simulation year. This file, called the "Marriage Pool", will be used in the second phase of the Demographic Block. The third output file contains the output records of all remaining individuals (all those who have neither died, emigrated, or been deemed ready for marriage). In phase one of the Demographic Block, then, all individuals who have not died or emigrated will have had all their demographic variables completely updated. Persons who will marry, on the other hand, are updated in all respects except for "pairing" and their province of residence.

The second phase is exclusively concerned with persons in the marriage pool and determines "who will marry whom". The individuals who had previously been recorded in the marriage pool (i.e., the second file mentioned above) are formed into couples on the basis of age, province of residence and education. The demographic characteristics of these couples, with the exception of province of residence, have all been updated in phase one and phase two now determines the province of residence of the new couples, in the same manner as is true of persons who are not marrying between  $t$  and  $t+1$ . (See section 4.1.3 below). All the couples in this "Marriage Pool" file are then merged with the records in the third output file (i.e., individuals who do not die, emigrate or marry during the simulation period) of phase one.



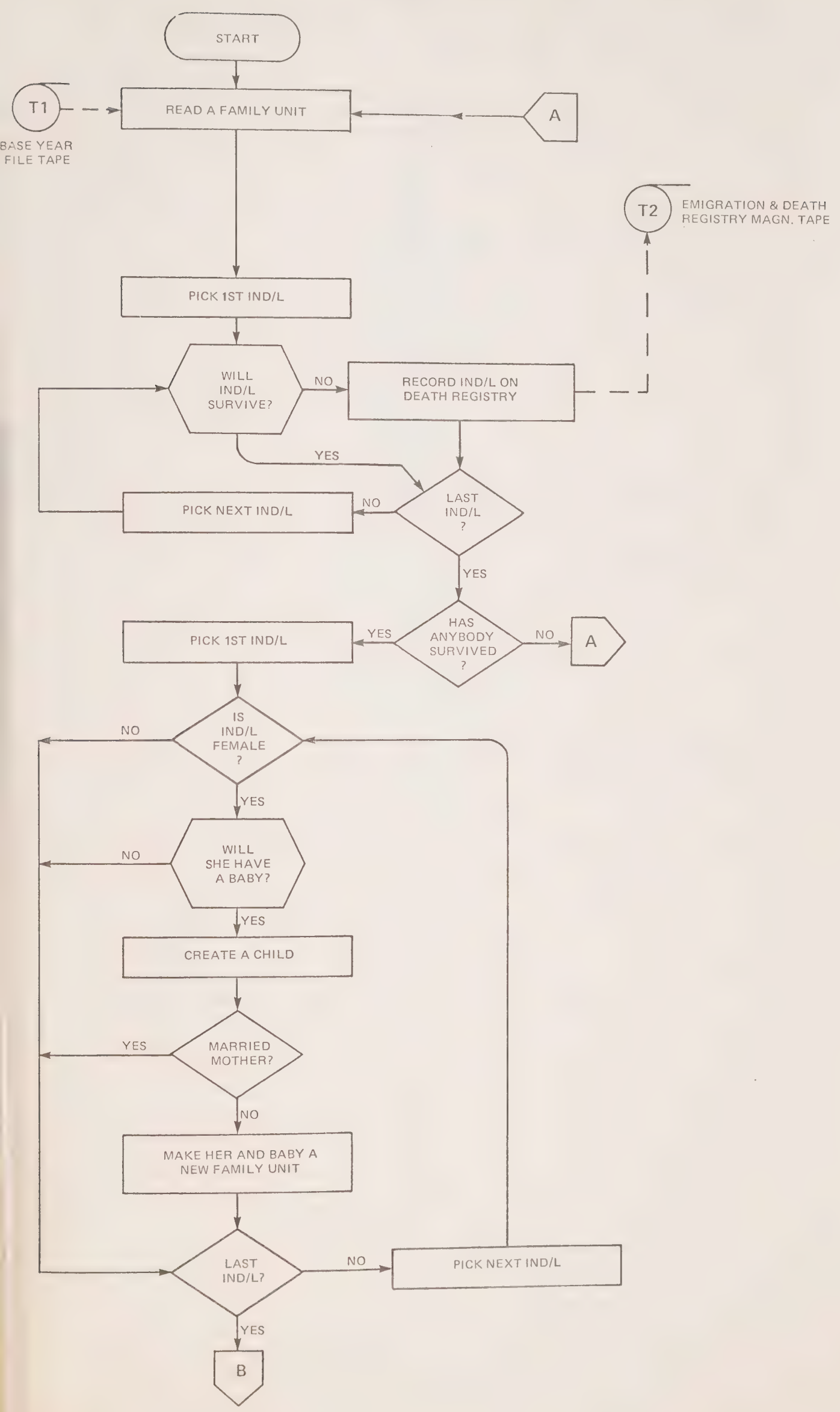
The basic method by which the individual state variables are updated is Monte Carlo simulation. Conceptually, this method is quite simple. Consider a married woman of age 17. The probability that she will give birth to a child within a period of a year is .45. If we wish to determine whether or not this woman will give birth to a child during a given year, we proceed as follows. We first choose a random number in the interval  $(0, 1)$ . If this number is less than or equal to .45, we decide that this particular woman will have a child during the simulated year. If the random number is greater than .45, she will not. All of the stochastic decisions in the POLSIM model are made in an analagous manner.

The logical structure of the Demographic Block can be seen in more detail with reference to the flow chart in figures 4.1 and 4.2. All decision processes that are to be resolved stochastically by the above Monte Carlo method are represented in the flow chart by hexagons. Deterministic decisions are indicated by the standard diamond branches.

The model begins by reading a family unit from the Initial Year Tape. The first decision to be made is that of whether or not the individuals in the family will survive. Survival is assumed to be an individual matter. All individuals within the given family are processed one by one. If the Monte Carlo procedure determines that the person will die, he is recorded on



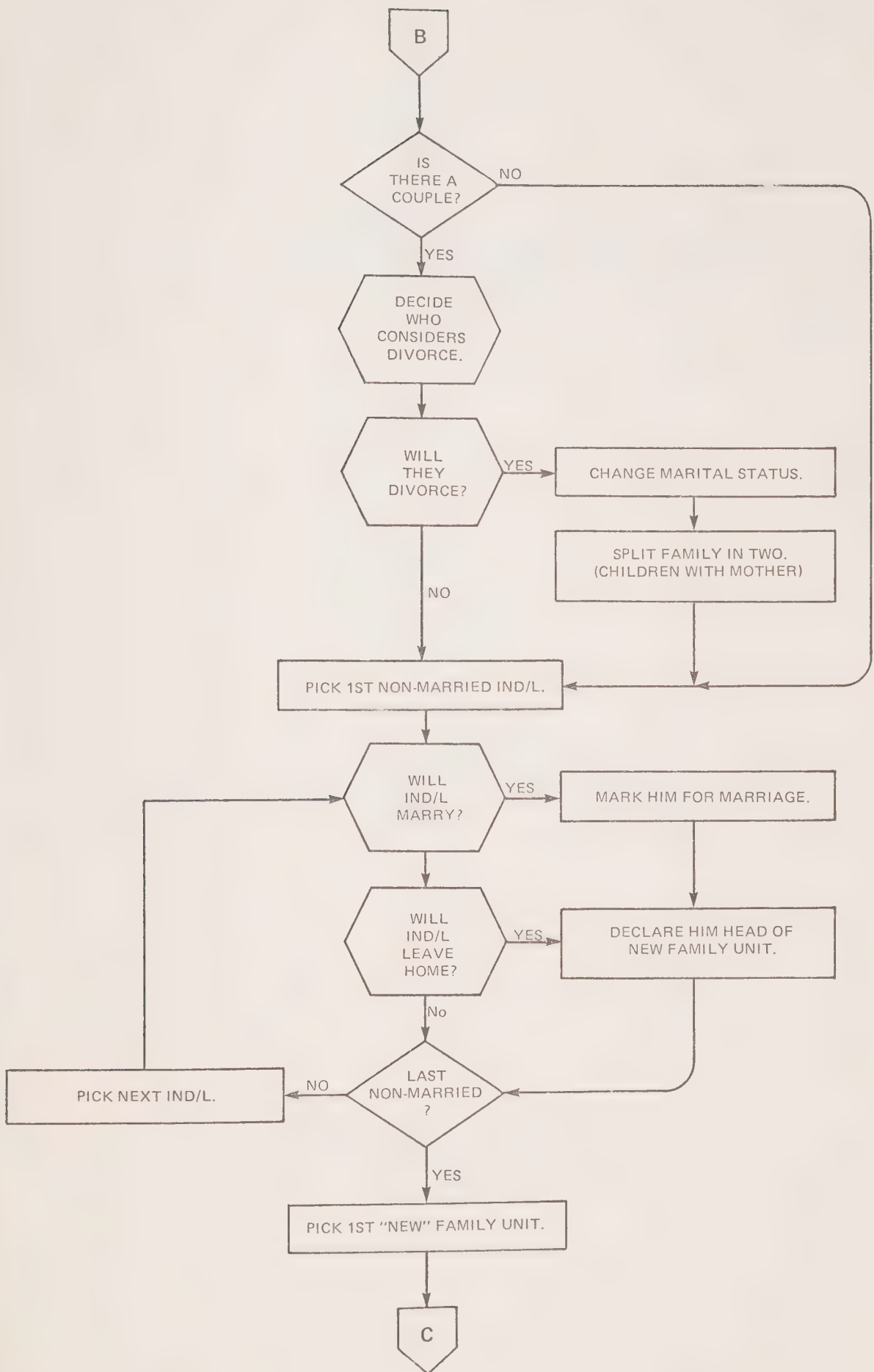
FIGURE 4.1 - DEMOGRAPHIC BLOCK FLOW CHART  
(PHASE 1)



(CONTINUED)









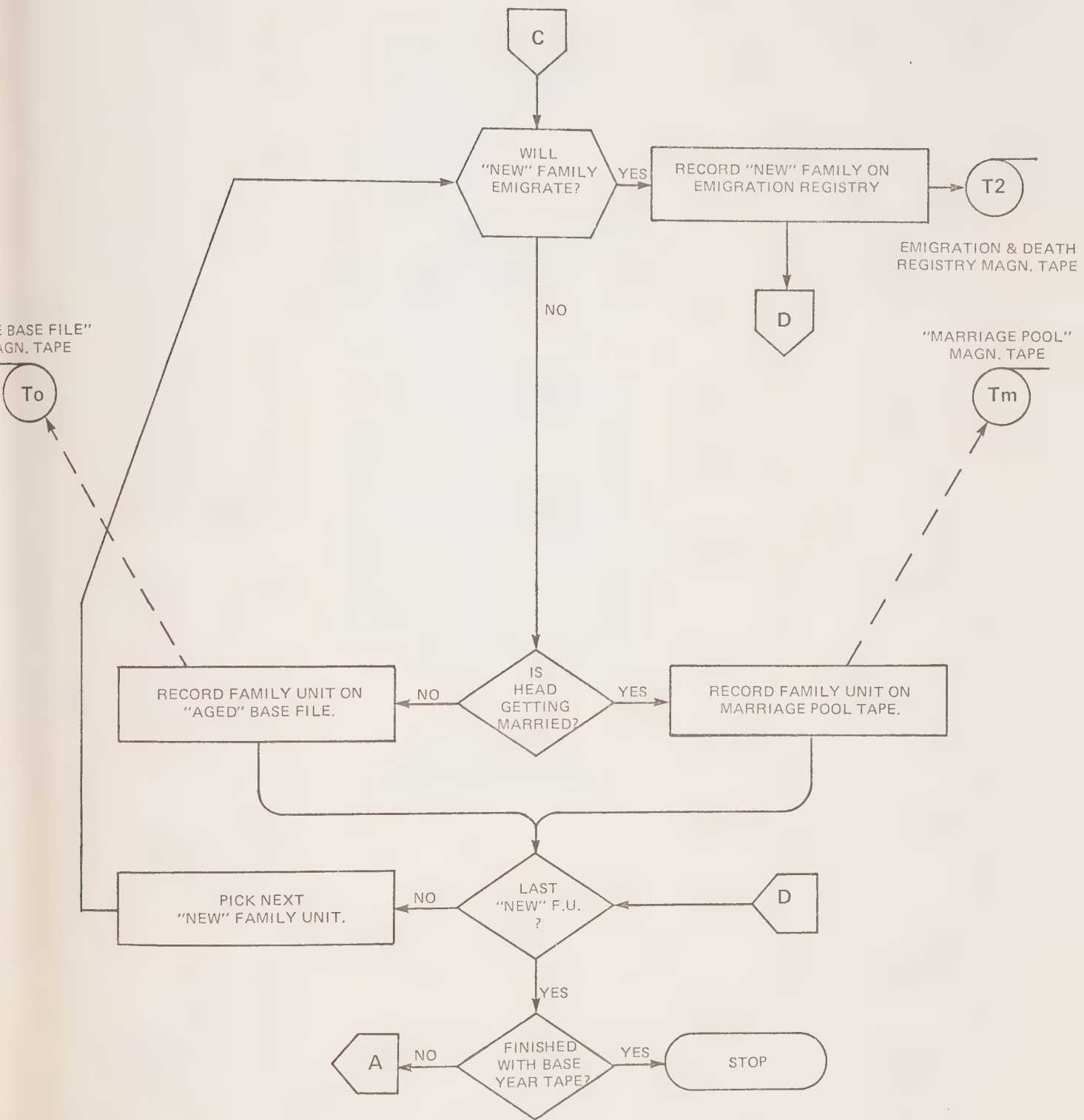
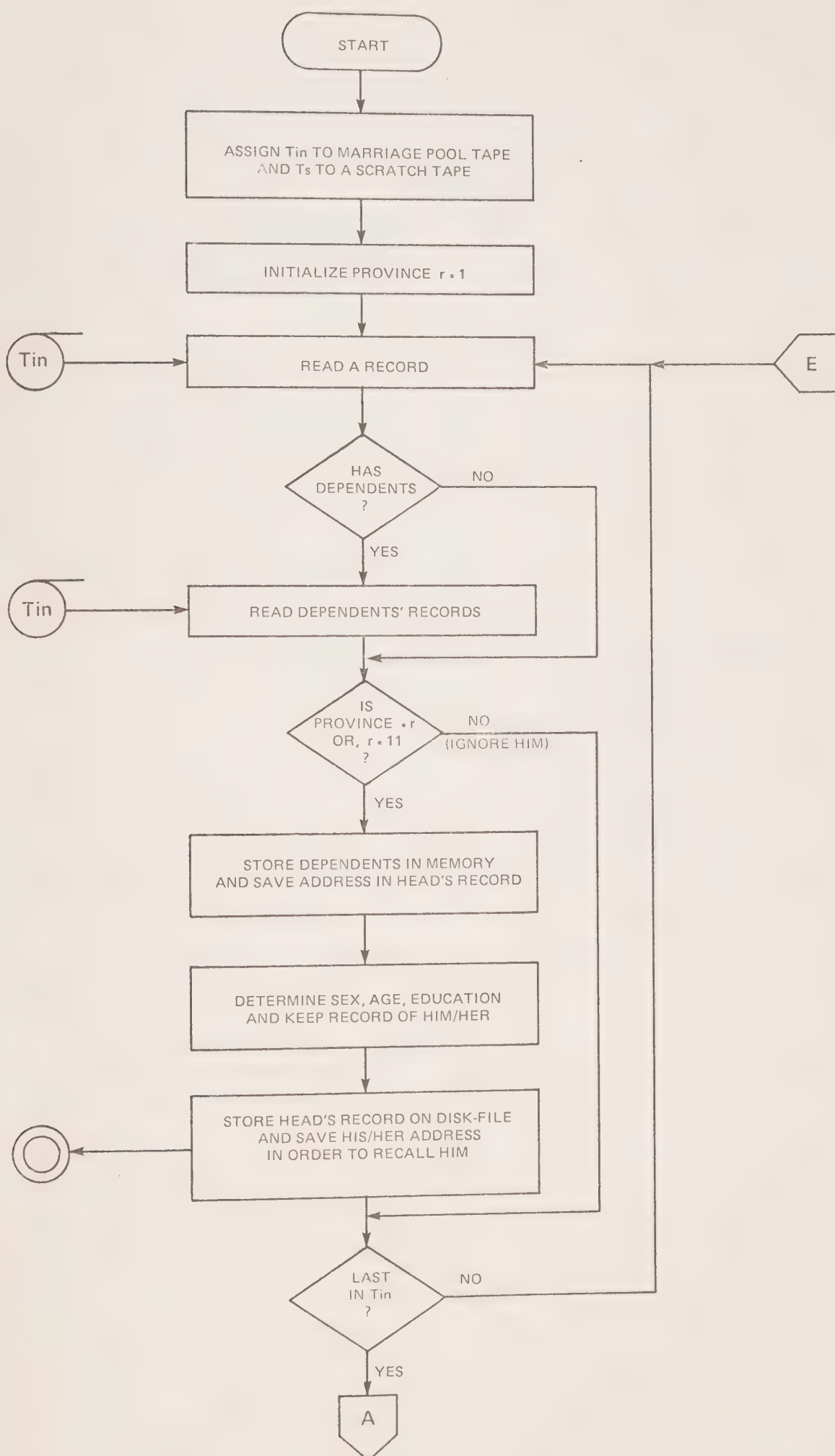


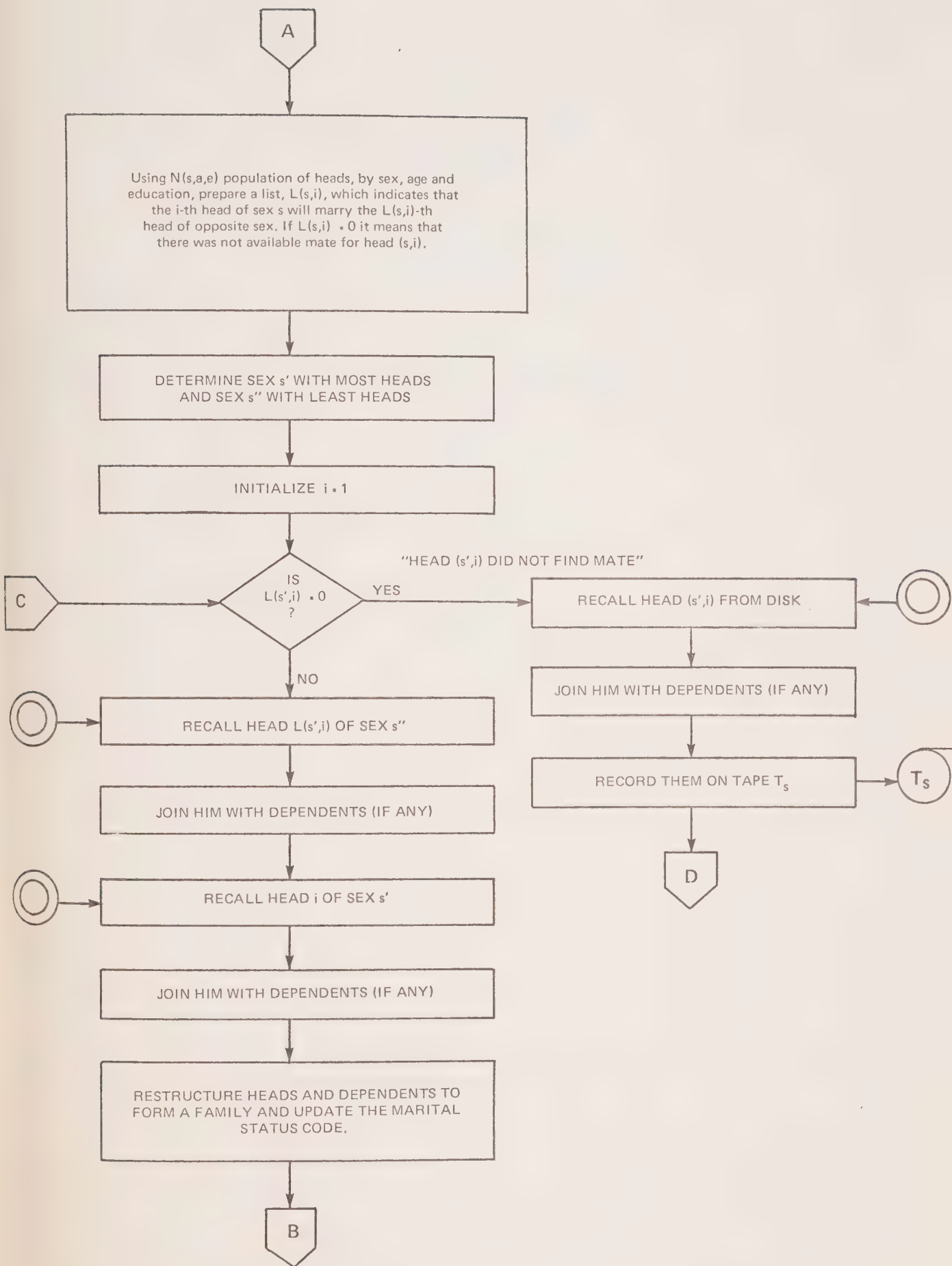


FIGURE 4.2 -- DEMOGRAPHIC BLOCK FLOW CHART  
(PHASE 2)

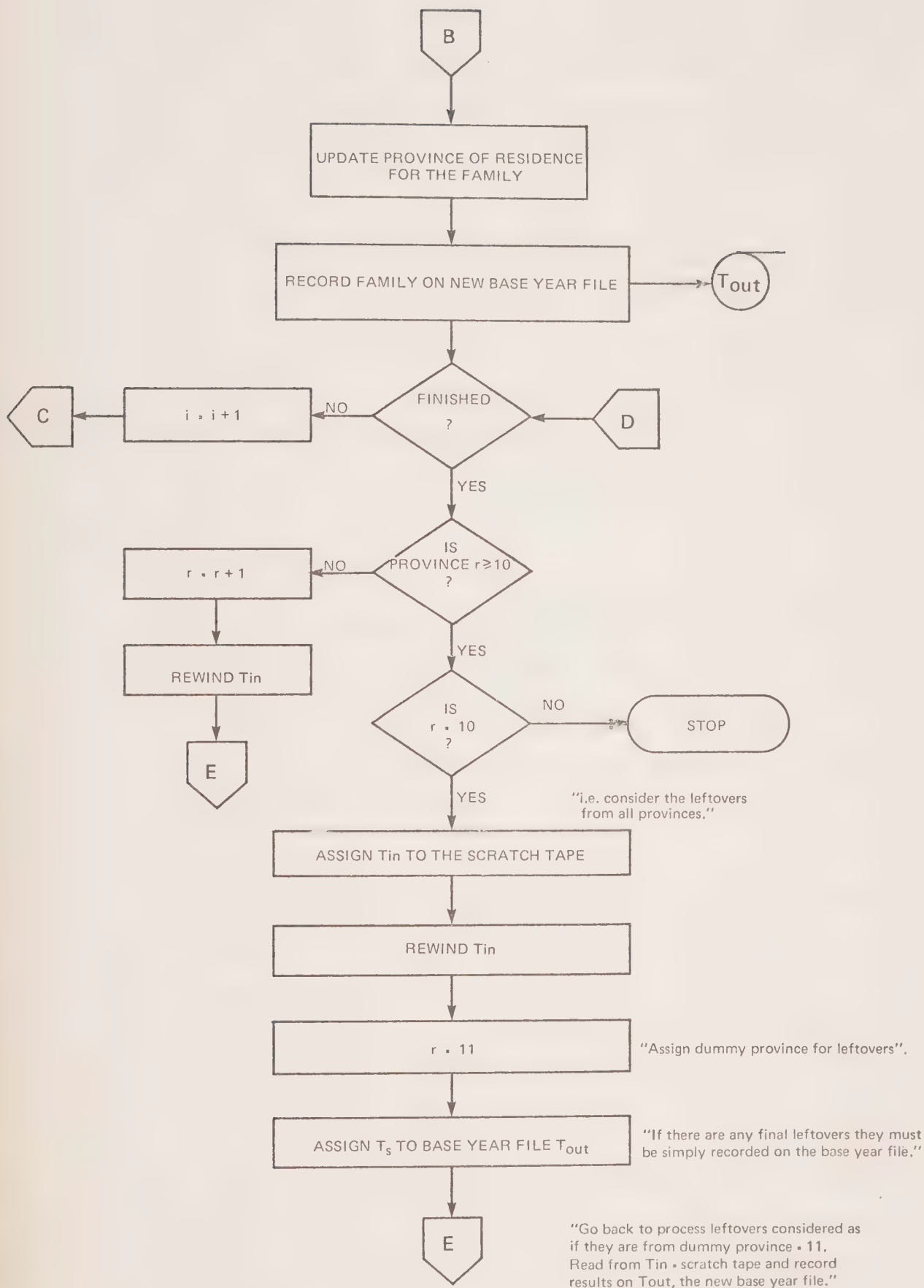














the death registry, and his record is not considered further. If nobody in the family survives, another family record is read from the Initial Year file. Persons who survive are aged by one year and then proceed to the birth process.

Births may be legitimate or illegitimate. If the woman is married, and it is determined that she will have a child, the child is simply added to the given family unit. If, however, she is not married and she is not the head of a family unit, she and her baby are assumed to form a new family unit; and, of course, she becomes the head of this new unit.

The divorce process is now commenced for all families containing both a head and a spouse. If divorce is determined to occur, the family is split in two. All children are assumed to go with the mother into a new, separate family unit with the mother as head.

All non-married individuals, excluding those who have become divorced in the current simulation, are now tested to determine whether they will get married. If this event occurs, the person is declared to be the head of a new family unit.

All individuals in each of the original families who are dependent (i.e., neither heads nor wives) are now tested to see whether they will leave home or not. If this event occurs, the individual is declared to be the head of a new family consisting of himself alone. It is clear that each



of the original family units might generate a whole set of new family units through marriage of dependents, divorce, birth of an illegitimate child, or as a consequence of a dependent leaving home.

For all families it is necessary to determine whether or not they will emigrate. This is considered to be a family decision, rather than an individual decision. All families that do emigrate are recorded on the emigration registry, and are not considered further.

Those family units that do not emigrate, with the exception of those who are going to be married in the current simulation, are now processed to determine their new province of residence. As mentioned above, province of residence of newly married couples is determined in phase two, after marriage has taken place.

The final step of the first phase is to record all family units on the proper files. Those families whose head is to be married are recorded on the marriage pool tape (file #2), while all others are recorded on file #3. (It will be recalled that file #1 contains the death and emigration registries). At this stage file #3 contains all remaining families, with the exception of new families formed through marriage. All demographic variables of the individuals on the third file have been completely updated.





In the second phase of the Demographic Block those individuals who are to be married are processed to form couples. Once this is completed, all the new husbands and wives, together with any dependents they might have, are formed into family units. The province of residence of these families is then determined, and they are then merged with the families of file #3 to create the completely updated initial year file. This is the final product of the Demographic Block, which now becomes input to the Activity Block that follows.

#### 4.1.3 The Demographic Block Processes

The previous section made reference to several "processes": birth, survival, marriage, etc. In this section we will elaborate on each of these. Before doing so, however, it will be worthwhile to briefly discuss the concept of a "process" itself. The term is very ambiguous, with a different meaning in almost every discipline. In mathematics, however, a process consists of a variable which is a function of time, and whose evolution over time is governed by certain underlying rules which might be either stochastic, deterministic, or some combination of the two.

Consider now the "death process" which we have referred to in the previous section. In light of the above definition, how is this process to be understood? All individuals can be assumed to exist in either one of two states: an individual can either be alive, or he can be dead. This suggests that we can define a "life state" variable,  $\ell(t)$ , which is defined in the following way:



$$\ell(t) = \begin{cases} 1 & \text{If he is alive at time } t \\ 0 & \text{otherwise} \end{cases}$$

Defined in this way, the variable  $\ell(t)$  is a stochastic process. The time of death, which is what we must determine, is postulated to be random. We attach to every individual a probability  $P$  that he/she will die in a period commencing at time  $t_0$  and terminating at time  $t_0 + \Delta t$ , where the interval  $\Delta t$  is fixed. In the POLSIM model, the time interval is one year.

In what follows we do not rigorously define each process in the above mathematical fashion. We simply describe the processes of the Demographic Block in a very general way. It should be understood, however, that underlying each process is an involved set of mathematical relationships: there is an ensemble of time functions, a state space, and a function mapping each element of the ensemble at each point in time to a unique element of the state space. There is also a probability distribution over the elements of the state space. This ensemble of time functions represents the histories of individual members of the nation's population, and are not, of course, completely known. What is known is a cross section of this ensemble at a given time and probabilities implicitly describing the possibilities of the future cross sections. The microsimulation technique selects from these possibilities in a random way, on the assumption that at the aggregate level the cross section properties will be preserved.

These relationships will only be implicit in the discussions that follow.



(a) The Death Process

This is a stochastic process applied universally to all individuals. The descriptive probabilities are dependent on age, sex, and time. These probabilities could be regionalized as well, but current evidence suggests that regional differences are too slight to warrant the added complexity. Full details as to the estimation and evaluation of the death probabilities are given in section 4.2 below.

(b) The Emigration Process

This is a stochastic process which is applied to family units. We postulate that whole families, rather than individuals, are the units that emigrate. The descriptive probabilities depend on the characteristics of the head of the family, namely, his marital status, age, and sex. These probabilities are assumed to be stationary in the present version of the model.

(c) The Birth Process

This is a stochastic process that is applied to females between the ages of 14 and 49. The descriptive probabilities depend on the age of the potential mother, her marital status, and her birth parity (i.e., the number of children borne alive to date).

(d) The Divorce Process

Divorce is a stochastic process that applies to married couples. It takes place in two stages. In the





first stage the spouse who will be the possible initiator of the divorce is determined. That is, it is assumed that if a divorce is to take place at all, one of the spouses, and only one, must initiate the process. It is assumed that it is equally likely that this will be the husband or the wife.

Once the possible initiator has been determined, the question of whether there will actually be a divorce is addressed. The probability of divorce depends now upon the sex and age of the initiating spouse. The decision is then made by the Monte Carlo procedure outlined earlier.

(e) The Marriage Process

Marriage is a stochastic process applied to individuals age 14 and over. It is decomposed into two sequential decisions. The first is the decision as to whether the person will get married or not. This is determined through a simple Monte Carlo procedure, where the descriptive probabilities depend upon the age, sex, region of residence, and marital status of the individual involved. The second decision deals with the problem of "who will marry whom". All individuals for whom it has been decided that marriage will occur are collected in a marriage pool. The second stage of the marriage process then matches the males and females in this pool.



The matching process proceeds as follows. The males and females are first partitioned into 300 age-education-province classes (ten provinces, ten age classes, and 3 education classes). Within each province, it is then necessary to designate a "choosing sex" and an "accepting sex". This is required because the descriptive probabilities for this process depend on sex. That is, we know the probability that a person of age " $a^1$ ", sex " $s^1$ " and education " $e^1$ ", will marry a person of age " $a$ " and education " $e$ ". Since we do not wish to give priority to either sex, we alternate which sex is the "choosing" and which the "accepting". More precisely, the algorithm adopts the following technique:

- (i) Begin by designating males as the "choosing sex" and females as the "accepting sex".
- (ii) Then sweep all of the age-sex classes in the choosing sex set, and within each set select approximately 10% of the population. Each of these selected individuals is then matched to an individual in the "accepting" set, on the basis of the above mentioned probabilities.
- (iii) Next switch the "choosing" sex. If there are more couples to be formed, we proceed as in step (ii). If not, the process simply terminates.

At this point we should mention that the matching probabilities are adjusted each time a certain sex,



age, education group becomes empty. It is clear that as matching progresses this is happening to the groups of individuals with same sex, age and education. When this happens for a certain group, it is evident that the probabilities for an individual of the opposite sex to marry somebody from the group in question must be adjusted to zero. On the other hand, the rest of the probabilities must be adjusted so that they represent probability distribution functions, i.e., they sum up to unity. The adjustment we make is proportional.

Since it is extremely unlikely that there will be an equal number of males and females designated for marriage within a given province, it is necessary to marry some individuals from different provinces. This will still likely leave an excess of one sex or the other, since the Canada wide totals are unlikely to be equal either. All individuals who are not able to find a mate in the given year are then recorded on the updated file with their original marital status. It was originally intended to give these individuals priority for marriage in the subsequent year, but the small size of this "leftover" population does not warrant the added complexity that would be involved.

(f) The Family Independency Process

This stochastic process is applied to all individuals who are dependents. The object of the process is to determine whether or not a dependent individual will leave his family or continue to stay within it for another year. The descriptive probabilities are dependent



on sex and age. In general, the probability of leaving home increases with age, then stabilizes and finally remains constant for all older ages, although it depends on sex.

(g) Interprovincial Migration Process

Like the emigration process, internal migration applies to family units, rather than individuals. The object of the process is to determine whether or not a family will move to another province, and if so, which one. First, one determines whether or not the family will remain in their present region. By region we mean the standard geographic division of Canada: the Atlantic provinces, P.Q., Ontario, Prairies and B.C. (Incidentally, more than 90% of the families stay in their region of residence.) The descriptive probabilities for this decision depend on the region, age, and income of the family head. Second, if the family moves out of its current region of residence, their new region of residence is determined. The descriptive probabilities of this decision form a transition matrix (with diagonal elements zero) which provides transition probabilities from region to region. This transition is postulated to be independent of income and age. However, one should notice that the decision to move out of a region depends on age and income of the family head. Third, if the new region of residence is P.Q., Ontario or B.C. the new province of residence is determined and no further action is necessary. If, however, the new region is either Maritimes or Prairies a decision is taken as to which specific province the family will move. The descriptive probabilities of this decision





form rectangular transition matrices (one for Maritimes and one for Prairies) which provide transition probabilities from (old) province of residence to the restricted set of provinces (either Atlantic or Prairies). Again this transition is postulated to be independent of incomes and age.

We assume that interprovincial migration of family units depend on the place of residence of the family head, his (or her) age and income. Statistical analysis showed that indeed all three characteristics are relevant to the probabilities in question. However, since the number of migrants is very small, compared to those who stay at their present place of residence, it was technically impossible to disaggregate transition probabilities by income and age. Instead we used the characteristics of age and income only to determine whether or not a particular individual will stay at his present region of residence. For those who move out of their region, their new residence is affected solely by their present place of residence.

#### 4.2 The Demographic Block Parameters

All of the processes of the Demographic Block, as described in section 4.1.3 above, entail certain descriptive probabilities. This section discusses the sources of these data, and the way in which they were estimated. The actual parameters are listed in Appendix C.



The probability parameters for all processes are assumed to be time-invariant with the exception of those for death. The parameters for birth and divorce are used directly as obtained from L. Stone of Statistics Canada. The parameters for emigration and marriage phase-I were estimated from raw data and care has been taken to be unbiased and consistent according to the sequential structure of the demographic block (see Appendix A.1). The death and family independence parameters were not estimated directly from raw data but by statistical inference. Finally, the parameters for marriage phase-II and interprovincial migration were computed in a straightforward way.

Almost all of the probability parameters were estimated from sources independent of our base year file. Their appropriateness for the model is checked in the validation section A.3 which again uses independent data for this purpose.

#### 4.2.1 The Death Process Parameters

The source of the Death Process parameters is Vital Statistics Division of Statistics Canada. The original data consisted of "Survival Ratios" for the years 1956, 61, 66, 69, 74, 79, and 84. A survival ratio is the number of surviving individuals in a given population at the end of a year divided by the total number of individuals in that population at the beginning of the year. These ratios can be identified as probabilities of survival (see Appendix A.1). The ratios are available by sex, and for single years of



age up to the age 100. Data for the first three years were historical estimates, while those for the last four were projections. The ratios are described in a Statistics Canada publication by W. Zayachkowskii.\*

As Zayachkowskii's paper suggests there is a time trend in these ratios. For young age groups this trend is upward and quite strong. For middle age groups there is practically no trend at all, and for older age groups the trend is weak and downward. We have developed a curve-fitting model in order to comprehend these time trends analytically. According to this model the survival ratio,  $S$ , for any sex and age is an exponential function of time:

$$S = c - \gamma \exp\{-\alpha(t-1970)\}$$

where,  $c$ ,  $\gamma$ ,  $\alpha$  are parameters dependent on sex and age.

There are three constraints imposed upon the parameters  $\alpha$ ,  $c$ , and  $\gamma$  as indicated below:

- (1)  $\alpha > 0$  because if  $\alpha < 0$  then the ratio  $S$  would become unbounded as  $t$  increases.
- (2)  $0 < c < 1$  because  $c = S$  in the limit as time tends to infinity. In other words, our time horizon is not limited by our model except insofar as future data might reflect revolutionary medical discoveries.
- (3)  $0 < c - \gamma < 1$  because  $c - \gamma = S$  for  $t = 1970$ .

---

\* W. Zayachdowskii, "Mortality Projections for the year 1969 DBS Population".





Our model is not typical of econometric or regression models. Special procedures were followed based on the fact that if we knew the parameter  $\alpha$  the model would become a simple linear regression.

Consider the following. Let three different points in time be  $t_1, t_2, t_3$ , and let the observed survival ratios be  $S_1, S_2, S_3$ , at these points respectively. If our model is going to fit exactly at these three points we will have

$$S_1 = c + \gamma \exp\{-\alpha(t_1-1970)\}$$

$$S_2 = c + \gamma \exp\{-\alpha(t_2-1970)\}$$

$$S_3 = c + \gamma \exp\{-\alpha(t_3-1970)\}$$

By subtraction we eliminate  $c$ , that is,

$$S_1 - S_2 = \gamma\{\exp\{-\alpha(t_1-1970)\} - \exp\{-\alpha(t_2-1970)\}\}$$

$$S_2 - S_3 = \gamma\{\exp\{-\alpha(t_2-1970)\} - \exp\{-\alpha(t_3-1970)\}\}$$

We apply the mean-value theorem of differential calculus by which,  $f(x_1) - f(x_2) = f'(z)(x_1 - x_2)$  where  $z$  is a number between  $x_1$  and  $x_2$ . In our case, we assume that

$$z = \frac{x_1 + x_2}{2} \quad \text{while, } f(x) = \exp\{x\}.$$

This gives

$$S_1 - S_2 = -\gamma\alpha(t_1 - t_2) \exp -\alpha \left( \frac{t_1 + t_2}{2} - 1970 \right)\}$$

$$S_2 - S_3 = -\gamma\alpha(t_2 - t_3) \exp -\alpha \left( \frac{t_2 + t_3}{2} - 1970 \right)\}$$



by dividing we have,

$$\frac{(S_1 - S_2)}{(S_2 - S_3)} = \frac{t_1 - t_2}{t_2 - t_3} \exp\{-\alpha(\frac{t_1 - t_3}{2})\}$$

Solving this equation with respect to  $\alpha$  gives

$$\alpha = \frac{2}{(t_3 - t_1)} \log\left\{\frac{(t_1 - t_2)}{(t_2 - t_3)} \frac{(S_2 - S_3)}{(S_1 - S_2)}\right\}$$

Therefore, in order that our curve passes through any given set of observations  $(t_1, S_1)$ ,  $(t_2, S_2)$ ,  $(t_3, S_3)$  the parameter  $\alpha$  must satisfy the above equation. Since we have seven observations which can be combined in groups of 3, i.e., 35 such combinations, we have  $\alpha_1, \alpha_2, \dots, \alpha_{35}$  estimates for the parameter  $\alpha$ . We take as a final estimate for  $\alpha$  the simple arithmetic average of these 35 estimates.

Finally, when the parameter  $\alpha$  is estimated our model becomes a simple linear regression:

$$S_t = c - \gamma \mu_t$$

where,  $\mu_t = \exp\{-\alpha(t-1970)\}$

The linear regression estimations of  $c$  and  $\gamma$  are reported in Appendix C for each age ( $a = 0, 1, \dots, 99$ ) and sex group.

#### 4.2.2 Emigration Process Parameters

Emigration probabilities were estimated from two sets of data. The first consisted of counts of total



emigrants broken down by marital status (married, non-married), sex, and 5 year age groups (0-14, 15-19, .... 70+).<sup>\*</sup> These counts were available for two periods, June 1968 - June 1969, and June 1969 - June 1970. The second set of data consisted of the Canadian population, partitioned into the same classes, for the years 1968, 1969 and 1970.<sup>\*\*</sup> Averages of populations for the years 1968 with 1969 formed the population at risk, while averages of emigrants for the two given periods formed the frequency of successes for the period January 1, 1969 to January 1, 1970. The ratio of emigrants divided by population gave us the estimates of the probabilities in question.

#### 4.2.3 The Birth Process Parameters

The birth process probabilities were supplied by L. Stone of Statistics Canada.<sup>\*\*\*</sup> These probabilities are stratified by age of the potential mother (14, 15, ...., 49), the number of children the mother already has (0, 1, 2, 3, 4, 5 or more) and on the legitimacy of the potential child.

---

\* Emigration totals were derived from unpublished data, Census Division, Statistics Canada.

\*\* Canadian population figures from Census Division, Statistics Canada.

\*\*\* Leroy O. Stone, "Preparation of Some Demographic and Socio-Economic Data Inputs", Statistics Canada Internal Report.



#### 4.2.4 Divorce Process Parameters

These were also provided by Leroy Stone.\* They are broken down by age and sex, and are listed in Appendix C.

#### 4.2.5 Marriage Process I Probabilities

These are the probabilities that a given individual will get married. They are available by region (Maritimes, P.Q., Ontario, Prairies, B.C.), by marital status (single, other), by five year age brackets, and by sex.

The marriage probabilities were derived from the following raw statistics:

- (i) The number of marriages that occurred in 1971 by age of bride, age of groom, marital status of bride (i.e. single, widowed, divorced, separated), and by province.\*\* It was assumed that the province where the marriage was registered was also the province of residence of both the bride and the groom. By proper aggregation we were then able to obtain the distribution of marriages by sex, age, marital status (reduced to single and "other" only) and region.

---

\* Leroy O. Stone, "Preparation of Some Demographic and Socio-Economic Data Inputs", Statistics Canada Internal Report.

\*\* Marriage statistics were obtained from unpublished data of the Vital Statistics Division of Statistics Canada.





- (ii) The population of unmarried individuals in 1971 by sex, age (5 year classes), marital status (single, other), and region.\*

The ratio of the first array to the second gave us the marriage ratios. These ratios, adjusted by death ratios (see Appendix A.1) gave us the probabilities in question. It should be noted that for the marital status "single" there were no recorded marriages for people 50 years and older. In other words, we had available only 7 age classes (15-19 to 45-49) for this group of people. It was therefore assumed that the probability of a single person over 50 getting married is zero.

#### 4.2.6 Marriage Process II Probabilities

From array (i) of the previous section we can obtain a cross classification of the number of persons who got married in 1971 by age of bride and age of groom. That is, we can determine  $N(a_H, a_W)$  which is the number of men in age group  $a_H$  that married women in age group  $a_W$ . We identify 10 age classes,\*\* and hence  $N$  is a  $10 \times 10$  matrix.

$$\text{Let } N(., a_W) = \sum_{x=1}^{10} N(x, a_W) \quad \text{sum of entries in column } a_W$$

---

\* Population statistics were obtained from Vital Statistics

\*\* The ten classes are: 14-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59, 60+.



$$\text{and } N(a_{H,.}) = \sum_{x=1}^{10} N(a_H, x) \quad \text{sum of entries in row } a_H$$

and for the matrices

$$M(a_H, a_W) = N(a_H, a_W) / N(a_H, .) \quad (10 \times 10 \text{ matrix})$$

$$\text{and } F(a_H, a_W) = N(a_H, a_W) / N(., a_W) \quad (10 \times 10 \text{ matrix})$$

Each element of M, say an element from row  $a_H$ , is obviously an estimate of the probability that a man in age group  $a_H$  will marry a woman in age group  $a_W$ . Similarly, each column of F is an estimate of the same probabilities for women.

These two matrices M and F can be combined into a 3 dimensional array  $P(s, a, a')$  which will be the probability that an individual of sex s who is in age bracket a will marry an individual of the opposite sex who is in age bracket  $a'$ .

We can derive a similar probability distribution relating the educational level of the two spouses. Specifically, we determine  $Q(s, e, e')$  which is the probability that an individual of sex s who is in education bracket e will marry an individual of the opposite sex who is in education bracket  $e'$ .<sup>\*</sup> There are three education classes: grade 8 and under, grade 9 to grade 13, and post-secondary. By assuming that the two random variables age and education are stochastically independent, we can calculate the joint distribution.

$$f(a, e, s | a', e') = P(s, a, a') \cdot Q(s, e, e')$$

---

\* This distribution was derived from the 1971 Survey of Consumer



which is the probability that a person of age  $a$ , education  $e$ , and sex  $s$  will marry an individual of the opposite sex of age  $a'$  and education  $e'$ .

#### 4.2.7 The Family Independence Parameters

The purpose of the family independence process is to declare dependent individuals independent of their family, and to cause make to them "leave home" on a probabilistic basis. Direct data for this process was not available, and it was therefore necessary to make inferences from available statistics. The estimation of the required probabilities was based on the model described below.

We assume that an individual may be in one of three states: (1) dependent, (2) independent or unattached, or (3) married, divorced, widowed, or separated. We will denote these states by  $d$  for dependent;  $i$  for independent, and  $m$  for "not-single". We ignore death or emigration, and transitions between the above 3 states are assumed to occur from year to year.

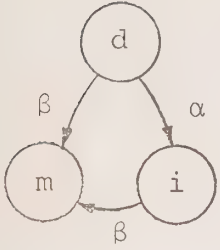
We make the following definitions:

$\alpha$	=	the probability that a person who is dependent in a given year will become independent in the subsequent year.
$\beta$	=	the probability that a person who is either dependent or independent in a given year will become "not-single" in the subsequent year.
$D_t(a)$	=	the population of $d$ type individuals at age $a$ at year $t$ .



$I_t(a)$  = the population of i type individual at age a at year t.

$M_t(a)$  = the population of m type individuals of age a at year t.



The shown diagram illustrates the transition from one state to another.

Since the system is closed we can easily derive the following identities:

$$I_{t+1}(a+1) = I_t(a) + \alpha D_t(a) - \beta I_t(a) \quad (1)$$

$$D_{t+1}(a+1) = D_t(a) - \alpha D_t(a) - \beta D_t(a) \quad (2)$$

$$M_{t+1}(a+1) = M_t(a) + \beta D_t(a) + \beta I_t(a) \quad (3)$$

By addition of terms in the above 3 equations we obtain:

$$I_{t+1}(a+1) + D_{t+1}(a+1) + M_{t+1}(a+1) = I_t(a) + D_t(a) + M_t(a) \quad (4)$$

which is as we would expect, since the system is closed and the population remains constant.

We further assume that the following relations hold:

$$\frac{I_{t+1}(a+1)}{I_{t+1}(a+1) + D_{t+1}(a+1) + M_{t+1}(a+1)} = \frac{I_t(a+1)}{I_t(a+1) + D_t(a+1) + M_t(a+1)} \quad (5)$$

$$\frac{D_{t+1}(a+1)}{I_{t+1}(a+1) + D_{t+1}(a+1) + M_{t+1}(a+1)} = \frac{D_t(a+1)}{I_t(a+1) + D_t(a+1) + M_t(a+1)} \quad (6)$$

If we add these two equations and subtract them from unity we obtain:

$$\frac{M_{t+1}(a+1)}{I_{t+1}(a+1) + D_{t+1}(a+1) + M_{t+1}(a+1)} = \frac{M_t(a+1)}{I_t(a+1) + D_t(a+1) + M_t(a+1)} \quad (7)$$





Relation (5), (6) and (7) assume certain stationary properties of the distribution of our population. For example, by (5) we can say that if the population of 20 year old independent individuals in 1968 was 7% of the whole population of 20 year old individuals this 7% is valid for the 20 year old population in 1967 or 1966 etc. The same interpretation can be given for relations (6) and (7) as well.

By substitution of (4) and (1) into (5) we obtain,

$$\frac{(1-\beta) I_t(a) + \alpha D_t(a)}{I_t(a) + D_t(a) + M_t(a)} = \frac{I_t(a+1)}{I_t(a+1) + D_t(a+1) + M_t(a+1)} \quad (8)$$

Similarly, substitution of (4) and (2) into (6) yields,

$$\frac{(1-\alpha-\beta) D_t(a)}{I_t(a) + D_t(a) + M_t(a)} = \frac{D_t(a+1)}{I_t(a+1) + D_t(a+1) + M_t(a+1)} \quad (9)$$

Equation (8) can take the form,

$$(1-\beta) + \alpha \cdot \frac{D_t(a)}{I_t(a)} = \frac{I_t(a+1)}{I_t(a)} \cdot \left( \frac{I_t(a) + D_t(a) + M_t(a)}{I_t(a+1) + D_t(a+1) + M_t(a+1)} \right) \quad (10)$$

and equation (9) can take the form,

$$1-\beta-\alpha = \frac{D_t(a+1)}{D_t(a)} \cdot \frac{I_t(a) + D_t(a) + M_t(a)}{I_t(a+1) + D_t(a+1) + M_t(a+1)} \quad (11)$$



By subtraction of (10) and (11) and solving with respect to  $\alpha$  we obtain,

$$\alpha = \frac{I_t(a)}{I_t(a)+D_t(a)} \cdot \frac{I_t(a)+D_t(a)+M_t(a)}{I_t(a+1)+D_t(a+1)+M_t(a+1)} \cdot \left( \frac{I_t(a+1)}{I_t(a)} - \frac{D_t(a+1)}{D_t(a)} \right) \quad (12)$$

Equation (12) is a formula providing an estimate of the probability that an individual will become independent. All of the terms on the right hand side can be derived from the Survey of Consumer Finance. The manipulation of the actual figures is presented in Appendix C, together with the other probability parameters. The model failed to give meaningful results for ages over 26 because the relevant populations became statistically insignificant. It was therefore necessary to assume a constant probability of independence for dependent individuals over age 26.

It should be noted that the "populations" referred to above pertain to a particular sex. The derived probabilities are therefore contingent on age and sex.

#### 4.2.8 Interprovincial Migration Process Parameters

The model of the internal migration process requires three sets of parameters, all of which were derived from a 5% longitudinal sample of tax filers collected by the Department of National Revenue for the years 1967-1970.\*

---

\* Department of National Revenue (Taxation), unpublished data.



The first set of parameters is the array  $q_1$  (R,I,A), which is the probability that an individual in Region R, income class I, and age class A\* will stay in his present region of residence. The second set is the array  $q_2$  (N,O), which is the conditional probability that a person who is leaving "old region" O will move to "new region" N. Note that since this array is conditional on the person moving somewhere else,  $q_2(x,x) = 0$  for all  $x = 1,2,3,4,5$ . The final set of parameters is required because if a person\*\* is found to move to either the Prairies or to the Maritimes, the actual province to which he is moving must be determined. Thus  $q_3(OP, NP)$  is the conditional probability that a person whose new region is either the Maritimes or the Prairies, and whose old province is OP (OP = 1, 2, ... 10) will move to the province NP (NP = 1, 2, 3, 4, 7, 8, 9). We note that this set of probabilities allows for migration within the Prairie provinces, and within the Atlantic provinces. For example, an individual who originally lived in Manitoba might be found through array  $q_1$  to not leave the prairie region. But array  $q_3$  might then establish that although he did not leave the prairie region, he did in fact move from Manitoba to Alberta.

#### 4.3 Validation of the Demographic Block

No simulation ever yields perfect results. The purpose of this section is to describe, in a general way, the reasons why deviations can arise between

---

\* The regions are Atlantic, P.Q., Ontario, Prairies, B.C. The income classes are 0-\$1,499; \$1,500-\$2,999; \$3,000-\$4,499; \$4,500-\$6,999 and \$7,500 and over.

\*\* Note then when we say "person" or "individual" we are really referring to a family head who is assumed to move his whole family with him.



simulated values and actual measured values. We then go on to analyze these errors in the context of the Demographic Block and to discuss the ways in which they can be eliminated or reduced.

#### 4.3.1 Additivity of Errors Principle

Consider the following:

1. Let the "population at risk" be of size  $N$ , and let  $p$  be the probability that a certain event will occur to any given individual in this population. For example, we might be considering a population of 100 males in Ontario, and the probability that any of these individuals will survive for the period of a year might be  $p = .9986$ .
2. If we now simulate the given event for the given population, we will achieve  $x$  "successes". In the above example, we might find that 90 of the original 100 males in Ontario survive. The expected number of successes is

$$E\{x\} = Np$$

We can write the actual number of successes as

$$x = Np + e_s$$

where  $e_s$  is the "absolute simulation error", which is expected to be zero, but will in fact be some positive or negative number.





3. Assume now that neither  $N$  or  $p$  are known precisely. That is, we know  $N'$  instead of the true  $N$ , and  $p'$  instead of the true  $p$ .

$$\text{Let } \Delta N = N' - N$$

$$\text{and } \Delta p = p' - p$$

The error  $\Delta N$  will be called "absolute initial population error", and the error  $\Delta p$  will be called "absolute parameter error".

4. Simulating a population of size  $N'$  with probability  $p'$  will yield  $\hat{x}$  successes, where

$$\begin{aligned}\hat{x} &= N'p' + e_s \\ &= (N + \Delta N)(p + \Delta p) + e_s \\ &= Np + N\Delta p + \Delta Np + \Delta N\Delta p + e_s\end{aligned}$$

From which,

$$\frac{\hat{x} - Np}{Np} = \frac{\Delta N}{N} + \frac{\Delta p}{p} + \frac{\Delta N\Delta p}{Np} + \frac{e_s}{Np}$$

If we ignore the second order term  $\frac{\Delta N\Delta p}{Np}$ , we can write

$$\frac{\hat{x} - Np}{Np} = \frac{\Delta N}{N} + \frac{\Delta p}{p} + \frac{e_s}{Np}$$

$$\text{or } \epsilon_{\text{total}} = \epsilon_N + \epsilon_p + \epsilon_s$$

which expresses the "additivity of errors principle".



We should note that  $\epsilon_{\text{total}}$  is the total error relative to the true expected number of successes. The errors  $\epsilon_N$  and  $\epsilon_p$  are expressed relative to the true population size and the true probability respectively. And finally,  $\epsilon_s$  is the simulation error relative to the expected number of successes.

We will now discuss each of these errors in turn, as they apply to the processes of the Demographic Block.

#### 4.3.2 Simulation Errors

Again consider a population of size  $N'$  subject to some event with probability  $p'$ . Let  $x$  be the number of "successes" for the group in question. As in the analysis of section 4.3.1, we know that the actual number of successes will be,

$$x = \bar{x} + e_s = N'p' + e_s$$

where  $e_s$  is the absolute simulation error. The simulation error will have expected value zero (if the random number generator is perfect), and its variance will be  $N'p'(1-p')$ . That is,

$$E\{e_s^2\} = \text{Var}\{x\} = E\{(x - \bar{x})^2\} = N'p'(1-p')$$

What we wish to do now is determine the expected bounds of this error. That is, we wish to set certain natural limits on the capability of micro-simulation.



We can establish confidence intervals for  $e_s^2$  by using Tshebysheff's Lemma.\* This proposition states that "for any non-negative random variable  $u$  with known expectancy  $\bar{u}$ , the event  $\{u < t^2 \bar{u}\}$  for any  $t > 1$  has probability greater than  $1 - 1/t^2$ ". Applying this lemma, we can set  $u = e_s^2$ ,  $\bar{u} = N'p'(1-p')$  and  $t = 2$ . This gives,

$$\text{Prob } \{e_s^2 \leq 4N'p'(1-p')\} > 0.75$$

or,

$$\text{Prob } \{|e_s| \leq 2\sqrt{N'p'(1-p')}\} > 0.75$$

Generally we are interested in the relative simulation error with respect to the expected number of successes. We therefore obtain,

$$\text{Prob } \left\{ \left| \frac{e_s}{\bar{x}} \right| \leq 2 \sqrt{\frac{1-p'}{N'p'}} \right\} > 0.75$$

We can state this in words. At a confidence level higher than 75% the relative simulation error  $\frac{e_s}{\bar{x}}$  with respect to the expected number of "successes" cannot exceed in magnitude the quantity  $2 \sqrt{(1-p')/N'p'}$ .

Table 4.1 gives the expectancy  $\bar{x} = N'p'$  for a binomial distribution. Table 4.2 contains the maximum relative simulation error (at higher than 75% confidence level) for selected values of  $N'$  and  $p'$ . By inspection of Table 4.2 we can see that for any given population size, the maximum relative simulation error is inversely and non-proportionately related to the size of the probability of the event. Similarly, for any given value for the probability of an event, the maximum relative simulation error is inversely and non-proportionately related to the size of the population at risk. Two conclusions can be drawn from Table 4.2.

---

\* Uspensky, Introduction to Mathematical Probability pp. 182-187.



TABLE 4.1

EXPECTED VALUES FOR BINOMIAL DISTRIBUTION

Probability of Event	Population Size															
	5	10	25	50	100	200	300	400	500	600	700	800	1000	1500	2000	
0.001	0.1	0.1	0.1	0.2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	1.0	1.5	2.0	
0.003	0.1	0.1	0.1	0.3	0.3	0.6	0.9	1.2	1.5	1.8	2.1	2.4	3.0	4.5	6.0	
0.005	0.1	0.1	0.2	0.3	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	5.0	7.5	10.0	
0.008	0.1	0.1	0.2	0.4	0.6	1.6	2.4	3.2	4.0	4.8	5.6	6.4	8.0	12.0	16.0	
0.016	0.1	0.1	0.3	0.5	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	10.0	15.0	20.0	
0.025	0.2	0.3	0.7	1.3	2.5	5.0	7.5	10.0	12.5	15.0	17.5	20.0	25.0	37.5	50.0	
0.050	0.3	0.5	1.3	2.5	5.0	10.0	15.0	20.0	25.0	30.0	35.0	40.0	50.0	75.0	100.0	
0.075	0.4	0.8	1.9	3.8	7.5	15.0	22.5	30.0	37.5	45.0	52.5	60.0	75.0	112.5	150.0	
0.100	0.5	1.0	2.5	5.0	10.0	20.0	30.0	40.0	50.0	60.0	70.0	80.0	100.0	150.0	200.0	
0.150	0.6	1.5	3.8	7.5	15.0	30.0	45.0	60.0	75.0	90.0	105.0	120.0	150.0	225.0	300.0	
0.200	1.0	2.0	5.0	10.0	20.0	40.0	60.0	80.0	100.0	120.0	140.0	160.0	200.0	300.0	400.0	
0.250	1.3	2.5	6.3	12.5	25.0	50.0	75.0	100.0	125.0	150.0	175.0	200.0	250.0	375.0	500.0	
0.300	1.5	3.0	7.5	15.0	30.0	60.0	90.0	120.0	150.0	180.0	210.0	240.0	300.0	450.0	600.0	
0.350	1.8	3.5	8.8	17.5	35.0	70.0	105.0	140.0	175.0	210.0	245.0	280.0	350.0	525.0	700.0	
0.400	2.0	4.0	10.0	20.0	40.0	80.0	120.0	160.0	200.0	240.0	280.0	320.0	400.0	600.0	800.0	
0.450	2.3	4.5	11.3	22.5	45.0	90.0	135.0	180.0	225.0	270.0	315.0	360.0	450.0	675.0	900.0	
0.500	2.5	5.0	12.5	25.0	50.0	100.0	150.0	200.0	250.0	300.0	350.0	400.0	500.0	750.0	1000.0	
0.550	2.8	5.5	13.8	27.5	55.0	110.0	165.0	220.0	275.0	330.0	385.0	440.0	550.0	825.0	1100.0	
0.600	3.0	6.0	15.0	30.0	60.0	120.0	180.0	240.0	300.0	360.0	420.0	480.0	600.0	900.0	1200.0	
0.650	3.3	6.5	16.3	32.5	65.0	130.0	195.0	260.0	325.0	390.0	455.0	520.0	650.0	975.0	1300.0	
0.700	3.5	7.0	17.5	35.0	70.0	140.0	210.0	280.0	350.0	420.0	490.0	560.0	700.0	1050.0	1400.0	
0.750	3.8	7.5	18.8	37.5	75.0	150.0	225.0	300.0	375.0	450.0	525.0	600.0	750.0	1125.0	1500.0	
0.800	4.0	8.0	20.0	40.0	80.0	160.0	240.0	320.0	400.0	480.0	560.0	640.0	800.0	1200.0	1600.0	
0.850	4.3	8.5	21.3	42.5	85.0	170.0	255.0	340.0	425.0	510.0	595.0	680.0	850.0	1275.0	1700.0	
0.900	4.5	9.0	22.5	45.0	90.0	180.0	270.0	350.0	450.0	540.0	630.0	720.0	900.0	1350.0	1800.0	
0.950	4.8	9.5	23.8	47.5	95.0	190.0	285.0	380.0	475.0	570.0	665.0	760.0	950.0	1425.0	1900.0	





## MAXIMUM RELATIVE SIMULATION ERROR\*

(Confidence Level Greater Than 75%)

Probability  
of:

Event	Population Size														
	5	10	25	50	100	200	300	400	500	600	700	800	1000	1500	2000
0.001	8.53	8.17	7.30	6.33	5.17	4.00	3.33	2.98	2.70	2.48	2.31	2.17	1.96	1.61	1.40
0.003	7.84	7.07	5.65	4.47	3.36	2.48	2.05	1.79	1.61	1.47	1.37	1.28	1.15	0.94	0.82
0.005	7.29	6.31	4.77	3.65	2.70	1.95	1.61	1.40	1.25	1.15	1.06	1.00	0.89	0.73	0.63
0.008	6.64	5.53	3.99	2.97	2.17	1.56	1.28	1.11	0.99	0.91	0.84	0.79	0.71	0.58	0.50
0.010	6.30	5.14	3.64	2.69	1.93	1.39	1.14	0.99	0.89	0.81	0.75	0.71	0.63	0.52	0.45
0.025	4.73	3.91	2.41	1.74	1.24	0.88	0.72	0.63	0.56	0.51	0.48	0.45	0.40	0.33	0.28
0.050	3.56	2.83	1.71	1.23	0.87	0.62	0.51	0.44	0.39	0.36	0.33	0.31	0.28	0.23	0.20
0.075	2.96	2.16	1.39	0.99	0.71	0.50	0.41	0.36	0.32	0.29	0.27	0.25	0.23	0.19	0.16
0.100	2.56	1.86	1.19	0.85	0.60	0.43	0.35	0.30	0.27	0.25	0.23	0.22	0.19	0.16	0.14
0.150	2.07	1.49	0.93	0.68	0.48	0.34	0.28	0.24	0.22	0.20	0.18	0.17	0.16	0.13	0.11
0.200	1.73	1.25	0.80	0.57	0.40	0.29	0.24	0.20	0.18	0.17	0.16	0.15	0.13	0.11	0.09
0.250	1.53	1.09	0.70	0.49	0.35	0.25	0.20	0.18	0.16	0.15	0.14	0.13	0.11	0.09	0.08
0.300	1.35	0.96	0.61	0.44	0.31	0.22	0.18	0.16	0.14	0.13	0.12	0.11	0.09	0.08	0.07
0.350	1.21	0.86	0.55	0.39	0.28	0.20	0.16	0.14	0.13	0.12	0.11	0.10	0.09	0.08	0.07
0.400	1.09	0.77	0.49	0.35	0.25	0.18	0.15	0.13	0.11	0.10	0.10	0.09	0.08	0.07	0.06
0.500	0.93	0.65	0.45	0.32	0.23	0.16	0.13	0.12	0.10	0.10	0.09	0.08	0.07	0.06	0.05
0.550	0.88	0.63	0.40	0.29	0.20	0.15	0.12	0.10	0.09	0.09	0.08	0.08	0.07	0.06	0.05
0.600	0.81	0.57	0.37	0.26	0.19	0.13	0.11	0.10	0.09	0.08	0.07	0.07	0.06	0.05	0.04
0.650	0.75	0.52	0.33	0.24	0.17	0.12	0.09	0.08	0.07	0.06	0.06	0.06	0.05	0.04	0.04
0.700	0.59	0.42	0.27	0.19	0.14	0.10	0.08	0.07	0.06	0.06	0.05	0.05	0.05	0.04	0.03
0.750	0.52	0.37	0.24	0.17	0.12	0.09	0.07	0.06	0.05	0.05	0.04	0.04	0.04	0.03	0.03
0.800	0.44	0.32	0.20	0.15	0.10	0.08	0.06	0.05	0.04	0.04	0.04	0.03	0.03	0.03	0.02
0.850	0.39	0.27	0.17	0.12	0.09	0.07	0.05	0.04	0.03	0.03	0.03	0.03	0.03	0.02	0.02
0.900	0.33	0.23	0.14	0.10	0.07	0.05	0.04	0.03	0.03	0.03	0.03	0.03	0.03	0.02	0.02
0.950	0.28	0.19	0.10	0.07	0.05	0.04	0.03	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02

\* Error relative to expected value.



First, small population sizes at risk ( $N'$ ) combined with small probabilities of success ( $p'$ ) will yield very large relative simulation errors. We can illustrate this with respect to survival simulations. The average probability of death for males in the 5-9 age bracket is  $p = .00063$ . The following Table 4.3 gives populations, expected number of deaths, simulated deaths, and the calculated simulation errors. The last row gives the maximum relative simulation error,

Table 4.3

Death Simulation: Males Aged 5-9 Years

Region	Maritimes	P.Q.	Ontario	Prairies	B.C.	Canada
Pop. at risk	2,536	7,808	8,195	3,849	2,207	24,595
Expected Value of deaths	1.59	4.91	5.16	2.42	1.39	15.49
Simulated deaths	2	6	5	5	0	18
Relative Simulation error	-25.8%	-22.2%	-1.2%	-106.6%	-100.0%	-16.20%
Max. relative simulation error (%)	159%	90%	89%	129%	170%	51%

From the above table it is clear that although our simulation results are acceptable, there is a very wide margin of simulation error that can arise when we are working with such unlikely events. In the case of Ontario, for example, the simulation was excellent. But it can be seen that this is purely the result of chance, in the strictest meaning of the word. We could just as easily have had a very large error. But it should be noted that relative error has little meaning



when we are considering very unlikely events. In these cases absolute error is the more relevant concept, and it can be seen that by this criterion the simulation performs very well.

It should be noted also that one of the reasons the relative simulation error seems very high is that it is defined with respect to the expected number of successes. Since the latter can sometimes be very small, the relative error can sometimes be very large. Alternatively, one could define "success" to be survival, rather than death. This will give us an indication of the extent to which the simulation error affects the whole population. This of course reverses the situation we had previously. The simulation survival error for B.C. (which is the province with the largest death relative simulation error) is now

$$\frac{1.39}{2207 \times .99937} = .063\%$$

and the maximum simulation error is

$$2\sqrt{\frac{.00063}{.99937 \times 2207}} = .106\%$$

We can thus see that the simulation error will create a deviation from the whole population by one thousandth at the most.



The second conclusion which we may draw (really the converse of the first) is that large population sizes at risk ( $N'$ ) combined with large probabilities of success ( $p'$ ) will yield small relative simulation errors. We will consider two examples.

- (a) The aggregate death probability is  $p = .007$ , while the total population in our 2% sample is approximately 400,000. The expected number of deaths is  $x = 2,800$  while the maximum relative simulation error could go as high as 3.76% (at a confidence level higher than 75%).
- (b) The aggregate fertility probability is  $p = .07$ , while the female population between the ages of 15 and 49 is  $N = 213,000$ . The expected aggregate number of births is  $x = 14,910$ , and the maximum relative simulation error could go as high as .78% (at a confidence level higher than 75%).

We can conclude, then, that at the national aggregate levels the birth and death processes will be almost entirely free of simulation errors.

The above analysis of simulation errors was based on error free populations at risk and probabilities. In practice, neither populations at risk nor probabilities are error-free. Therefore, exact evaluation of the simulation error is not possible. However, the maximum relative simulation error can still be evaluated, using the above suggested expression  $2\sqrt{(1-p)/(Np)}$ , at the confidence





level of 75%. Values of this expression are contained in Table 4.2. In appendix C we do not calculate these errors but they can be either calculated or obtained directly from Table 4.2.

#### 4.3.3 Initial Population Errors

According to the Additivity of Errors Principle (cf. Section 4.3.1) any imperfection of the initial year tape will generate errors for the simulated population of the following year. No action is taken to rectify this, but an assessment of these so-called initial population errors is always necessary. This necessity is obvious because these errors put a lower bound on the total errors. For example, if there is a 3% overestimate in the number of females of age 20-24 in the initial year file, then the simulated births from these women will, under perfect conditions (i.e., correct birth probabilities and negligible simulation error), be 3% too high. Of course, the total error might be higher, e.g., 7%, in which case the other two error components (simulation and parameter) had an additive effect. Or the total errors might be lower, e.g., 1% or, -2%, in which case the other two errors had cancelling effects.

The initial population errors have cumulative effects when the simulation is done over a period of more than one year. The present report does not attempt to assess the extent of this cumulative error. Instead, Appendix C.3



simply documents the size of these errors, as they pertain to certain of the Demographic Block processes. For example, from table C.31 we have for Quebec a 12% underestimate of the 65-69 years old females, while the same age bracket males in Ontario are underestimated by 18.5%. Also the males in Ontario aged 25-29 are overestimated by 9.2%; this is one of the few overestimates we have on the initial population as of April 1, 1968. These are some of the worst initial population errors. Most do not exceed  $\pm 5\%$  as one can verify from table C.31. Finally, one could observe from table C.32 that the base year file underestimates the number of married women 15-19 years of age by 24.2%. This indeed is a serious shortcoming because this group has high fertility probabilities and this particular underestimate contributes to initially large errors on the number of births. As the simulation proceeds, of course, the birth error is progressively corrected as more accurately estimated initial population female cohorts move into their high fertility years.

#### 4.3.4 Validation and Parameters' Calibration

The various probability parameters used in the Demographic Block are provided from various sources. Some of them were estimated from raw data by taking ratios of successes over populations at risk. Others were supplied ready for use from other studies. Finally some were estimated by inference from time series or from simplified models. It is evident that before we use the probabilities of the Demographic Block, they should be first validated and if necessary adjusted.



In principle the Validation problem can be stated as follows:

Let  $p$  be the probability that a certain event will take place to an individual with certain characteristics. Also let  $N$  be the observed population of such individuals and  $A$  the observed number of successes of the event in question. Under the hypothesis that  $p$  is the correct probability the confidence interval for the successes at a confidence level of 90% is the interval:

$$I = \{Np - 1.65\sqrt{Np(1-p)}, Np + 1.65\sqrt{Np(1-p)}\}$$

If the observed successes  $A$  are within this interval then there is no evidence against the correctness of the probability  $p$ . If, on the other hand, this is not the case then we assume that  $p$  is incorrect and that it should be adjusted and replaced by:

$$P_{\text{new}} = Cp$$

where, the correction factor  $C = \frac{A}{Np}$

The above procedure can be applied on aggregate statistics rather than the stratified ones which in general are not available. For example, the birth probabilities are provided by Stone in single years of age of the potential mother, her birth parity, and her marital status. For the year 1968, however, we had reliable statistics on population at risk and observed



births by marital status, i.e., legitimate and illegitimate cases, and 5 years age brackets. The birth parity stratification was not available. For this reason instead of the original probabilities we obtained average probabilities stratified by age and marital status only. The average of these probabilities is theoretically justified by Poisson's Theorem on weighted averages (see J.V. Uspensky - Introduction to Mathematical probabilities - pp. 208-215). On these average probabilities we applied the validation procedure outlined above. If a certain average probability needed to be corrected, then the corresponding correction factor was applied to the whole set of original probabilities whose average is the one in question.

In Appendix C.2 the validation results for the emigration, birth, marriage, divorce and survival processes are presented. The processes are validated and the probability parameters are calibrated in the above mentioned way. However, no validation was done on the Family Independence Parameters due to lack of complete information.

The Interprovincial Migration Parameters were validated in the following simple minded way, due to a lack of any information other than the original data from which these parameters were estimated. We ignored birth, immigration, emigration and death and analytically we estimated the distribution of population in April 1972 from the Consumer Finance distribution in April 1968.





This was done by moving families from province to province on a year to year basis. The process was analytical and implied the closed Markovian system identity:

$$x_{t+1} = Px_t$$

where  $x_t$ ,  $x_{t+1}$  are the family counts vector at year  $t$  and  $t+1$ , respectively, and  $P$  is the transition matrix which depends on income and age of the heads of the family units in question. In Appendix C these results are presented and compared with the distribution of population in April 1972 as recorded on the corresponding Consumer Finance Survey tape.



## 5. The Activity Status Block

This chapter is divided into five main sections. Section 5.1 gives a brief overview of the block as a whole. Section 5.2 describes the logic of the simulation in detail, and makes explicit all the various assumptions that are made. Section 5.3 discusses the labor force model that is the heart of the Activity Block, and details all the mathematical adjustments that were required to refine this basic model. Section 5.4 is concerned with indicating how well the model performs, as compared with historical data. Section 5.5 gives a complete description of all of the data that is used, and the sources from which it was obtained. Appendix D provides greater detail, lists the data and documents the relevant computer software.

### 5.1 General Overview

#### 5.1.1 Purpose of the Activity Block

Broadly speaking, the purpose of the Activity Status Block is twofold: first, to update those variables in the individual state vector that describe what a person is "doing" during the year being simulated; and, second, to make certain adjustments to the individual state vector preparatory to updating the person's income in the Market Income Block. Specifically, the Activity Block does the following: (a) it determines the number of weeks during the year being simulated that the person spends in school, employed, unemployed, and in the non-labor force (where "non-labor force" is defined so as to exclude those in school); (b) it determines whether a person advances his education; (c) it determines the person's



activity at the end of the year being simulated or the beginning of the next simulation year (the "year" being simulated is defined as April through March); (d) it makes any necessary changes in a person's employment category ("Type"); (e) and finally, it converts a person's annual wages to a weekly wage rate if he was employed as a Class B person in the year preceeding the one being simulated. (A "Class B" person is one who is subject to unemployment). The Activity Block makes the requisite changes in the individual's state vector on the basis of his present year's demographic characteristics, his past year's activities, and exogenously determined monthly Canadian aggregate rates of unemployment.

#### 5.1.2 Methodology and Data

The methodological approach taken by the Activity Block is somewhat different than that taken by the other blocks in POLSIM. In the Market Income Block, for example, the various components of an individual's income are simulated directly. A transition matrix determines to what extent, if any, a particular component of a person's income is to be increased or decreased during the simulated year. In the Activity Block such direct transitions are not possible. Instead, a person is thought of as being in one of four activity states: employment, unemployment, non-labor force, or school. Transitions among these four states then take place month by month, tracing out a year long "activity history" for a given individual. The number of weeks in each of the four states, any change in education status, and any changes in employment category are then inferred from this history.



Much of the Activity Block is concerned with the problem of calculating the transition matrices that are used to determine how a person will move from month to month. That is, these matrices are for the most part not read in as data, but rather are calculated from other data in the model. How this is done will be discussed in some detail in Section 5.3 below.

### 5.1.3 The Activity Variables

There are seven variables in the individual's state vector that the Activity Block is concerned with.

#### 1. Weeks in School

This is simply the number of weeks during the year (the year being defined as April through March) that the person spends in one or more of the 19 education states.

#### 2. Weeks Employed

The number of weeks during the year that the person spends in the employed state.

#### 3. Weeks Unemployed

The number of weeks during the year that the person spends in the unemployed state.





4. Weeks in non-labor force

The number of weeks during the year that the person spends in the (narrowly defined) non-labor force. The Activity Block distinguishes two kinds of non-labor force status. The generalized non-labor force (GLF) includes all people who are neither employed nor unemployed. Thus students, housewives, and children would fall into this category. The narrowly defined non-labor force (NLF) is the same as the GLF, except that it does not include students.

5. Education

There are 19 possible education states that a person might be in. They are self-explanatory:

1	Grade	9	11	Univ	3
2	Grade	10	12	Univ	4
3	Grade	11	13	Univ	5
4	Grade	12	14	Univ	6
5	Grade	13	15	Univ	7
6	CAAT	1	16	Univ	8
7	CAAT	2	17	Univ	9
8	CAAT	3	18	Univ	10
9	Univ	1	19	Less than Grade	9
10	Univ	2			

6. April Activity Status

The April Activity Status defines what a person is doing in April of the year following the one being simulated. The reason for this variable is that for the month-to-month simulation it is necessary to know what state the person will be in when he starts out.



April Activity Status thus provides continuity from year to year.

There are 23 Activity States, and they again are self-explanatory.

1	Grade	9	13	Univ	5
2	Grade	10	14	Univ	6
3	Grade	11	15	Univ	7
4	Grade	12	16	Univ	8
5	Grade	13	17	Univ	9
6	CAAT	1	18	Univ	10
7	CAAT	2	19	Unused	
8	CAAT	3	20	Employment	
9	Univ	1	21	Unemployment	
10	Univ	2	22	Non Labor Force	
11	Univ	3	23	Age less than 14	
12	Univ	4			

7. Employment Category (Type)

Different kinds of people relate to the labor force in different ways. The Activity Block distinguishes people who will never become unemployed, and people who will; people who are retired and people who are not; people who are in pensionable employment and people who are not; and people who have become employed for the first time or retired for the first time. The Employment Category is a sort of catch-all variable that supplies all of this information.

There are 13 possible values for "TYPE":

- (a) 14 - This is a "Class A" person (a male who is either self-employed, or employed in a professional, technical, or managerial capacity, and who by assumption may never become unemployed) who on retirement will receive no private pension.



- (b) 15 - A "Class A" person who on retirement will receive a private pension.
- (c) 24 - A "Class B" person (by definition not a "Class A" person; someone who is subject to unemployment) who on retirement will receive no private pension.
- (d) 25 - A "Class B" person who on retirement will receive a private pension.
- (e) 3 - A person who has never been a member of the labor force.
- (f) 4 - A retired person who does not receive a private pension.
- (g) 5 - A retired person who is eligible for private pension.
- (h) 140, 150, 240, 250, 40, 50.

These are exactly the same as 14, 15, 24, 25, 4 and 5 except that the person has not yet had an initial income assigned to him, or has just had an initial income assigned but has not yet gone through an income transition.

#### 5.1.4 General Organization of the Activity Block

A general picture of the Activity Block can be obtained from the Macro Flow Chart given in figure 5.1. The



# ACTIVITY STATUS BLOCK

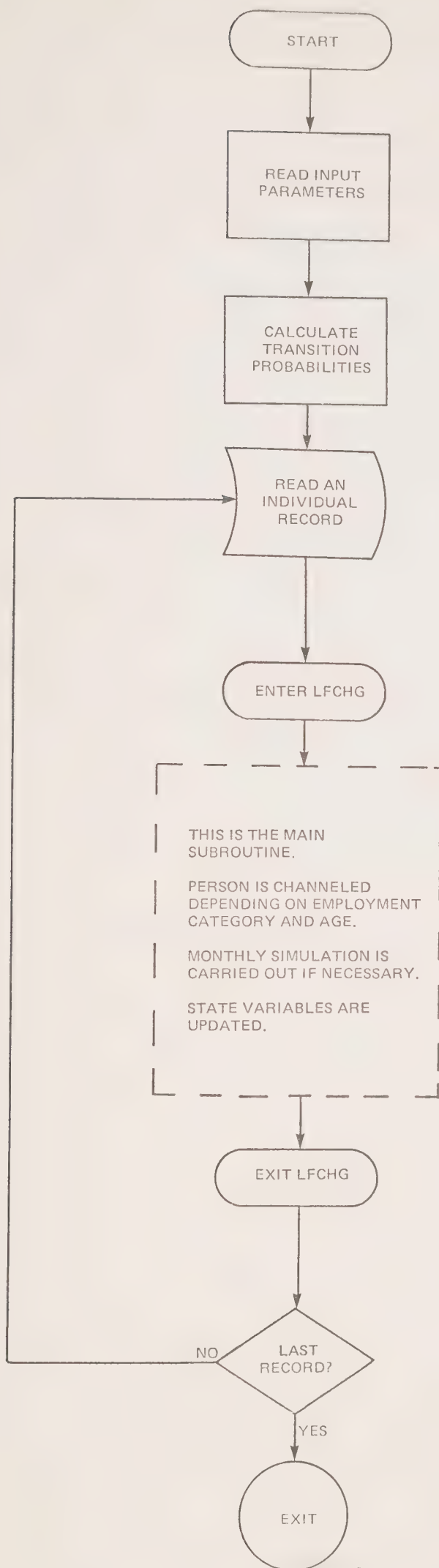


FIGURE 5.1 — MACRO FLOW CHART OF THE ACTIVITY STATUS BLOCK





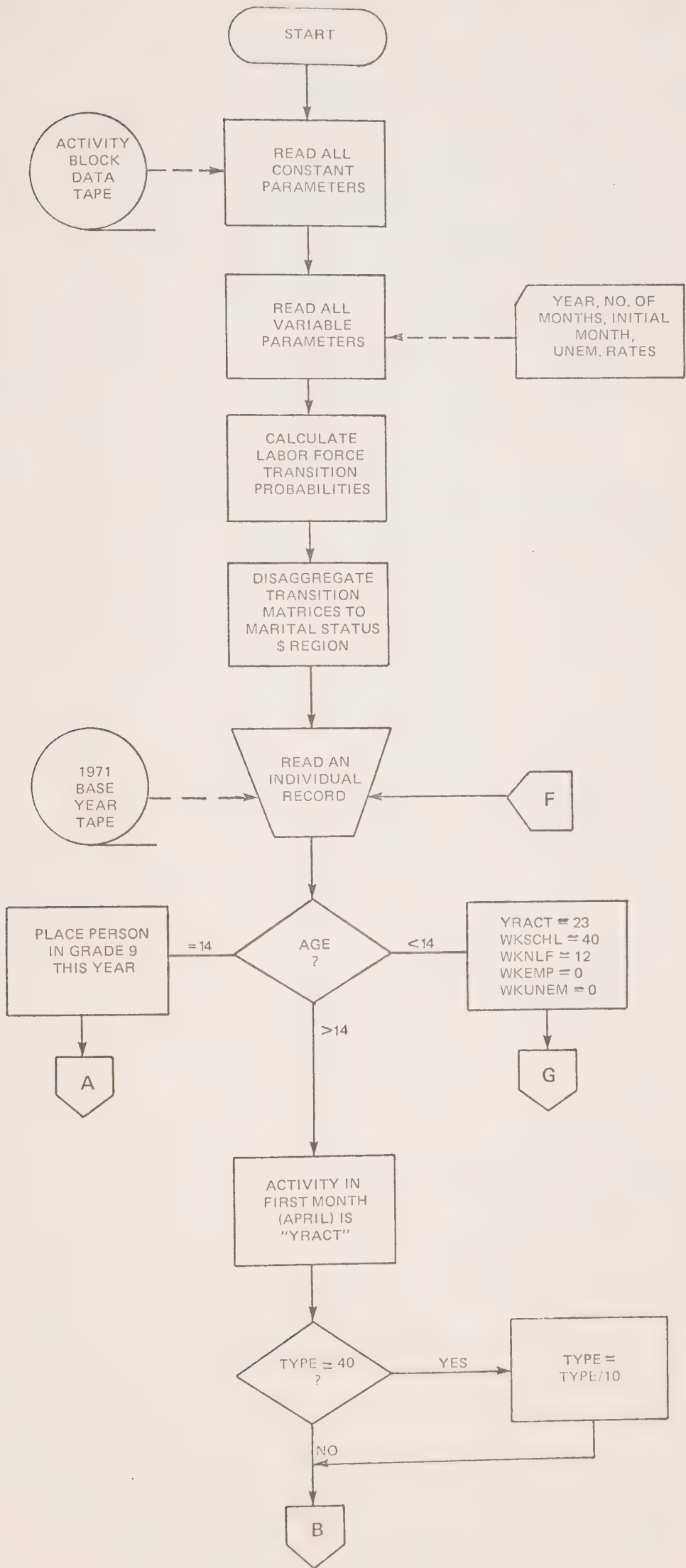
program begins by reading in all of the input parameters. These parameters include regression coefficients and unemployment rates from which an initial set of transition matrices are calculated, parameters that are used to adjust these probabilities in various ways, and probabilities used to determine whether and how a person advances through school. Once all of the data is properly constructed, individuals are passed through the logic of the simulation model proper. In this section persons are directed through various processes, depending on how they relate to the labor force (their Employment Category), and a few other factors. Children, for example, bypass most of the Activity processes completely. And persons in employment categories 14 or 15 bypass the monthly labor force simulation. The relevant state variables are then updated, and the whole procedure is repeated for the next individual.

## 5.2 Detailed Organization and Assumptions of the Activity Block

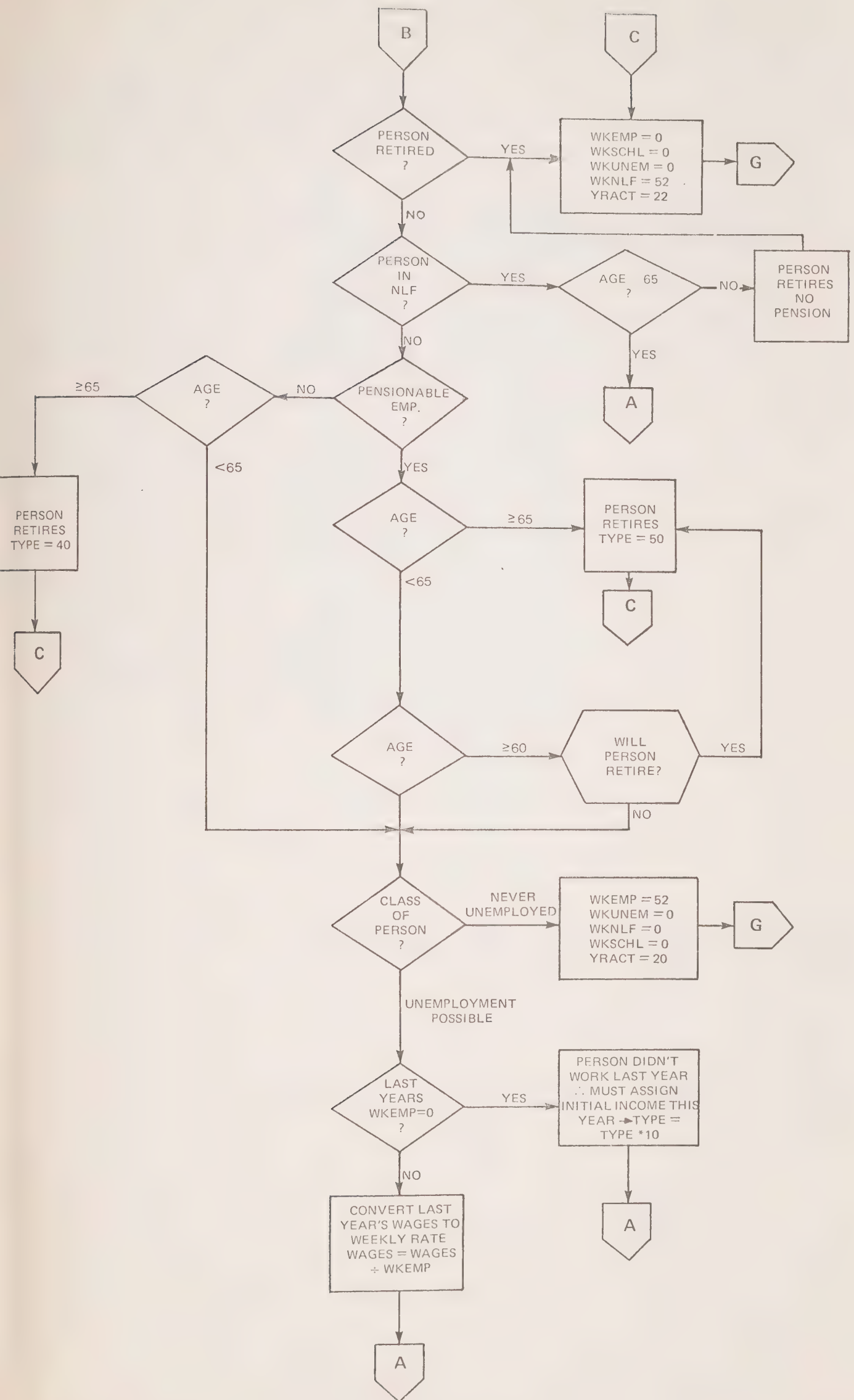
The analysis of an individual's activities can be traced with reference to the flow chart in figure 5.2. It defines the way the activity variables of an individual's state vector are updated. The idea behind the whole process is to infer what an individual does during a year by examining what he does in every month of that year. A year is assumed to run from April through April. (Thus providing a starting point for the next year's simulation.) An individual thus starts out in April doing something (being in a particular education state, being employed, etc.) and this then influences



FIGURE 5.2 – THE ACTIVITY BLOCK

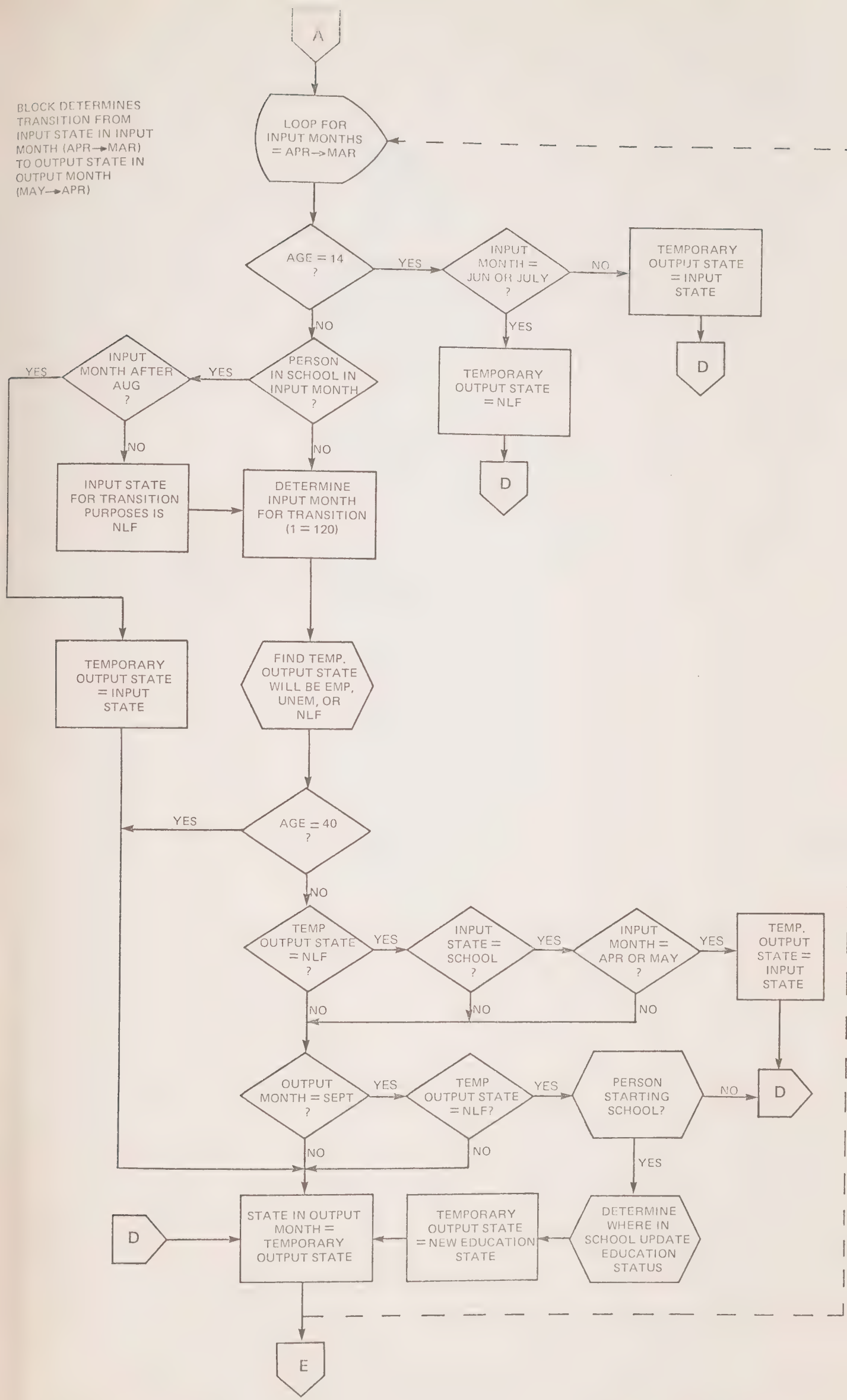






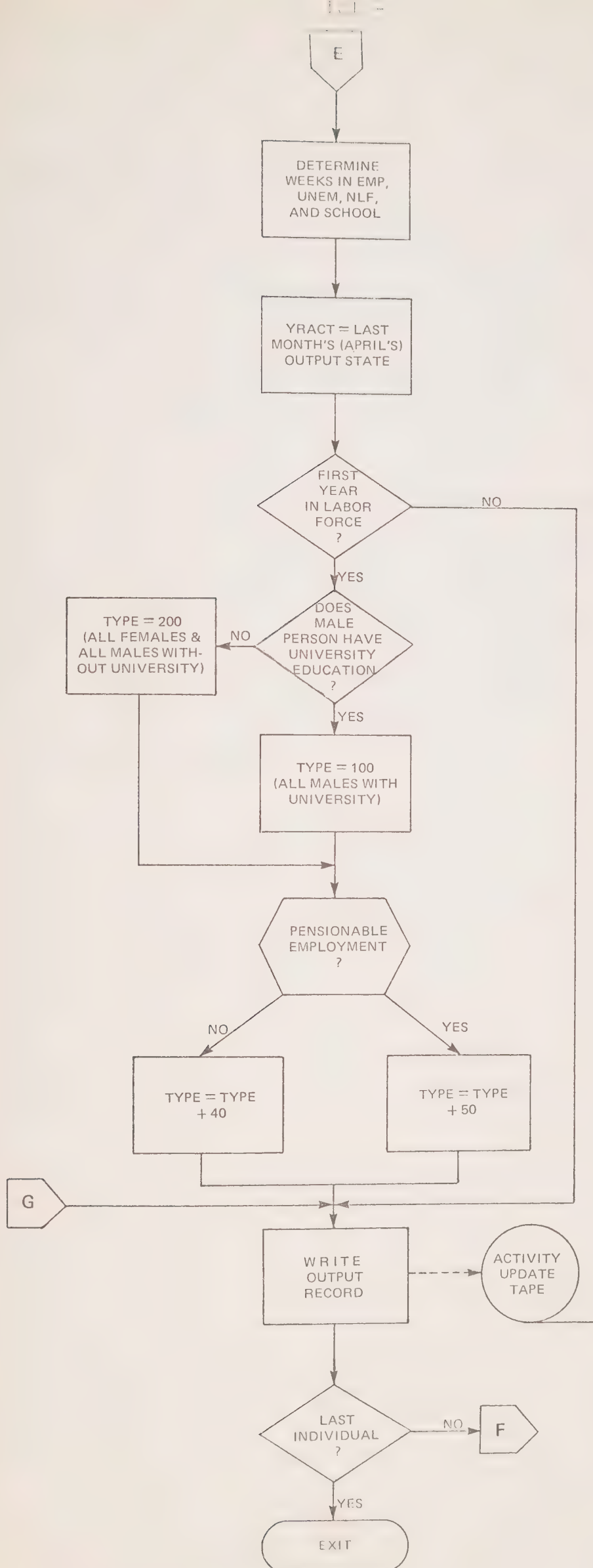


BLOCK DETERMINES  
TRANSITION FROM  
INPUT STATE IN INPUT  
MONTH (APR→MAR)  
TO OUTPUT STATE IN  
OUTPUT MONTH  
(MAY→APR)











(together with the exogenously inputted influence of aggregate unemployment) what he will be doing in May, and so on. The whole year is traced out in this way. An individual moves from state to state, depending on his present state, his age, the month being considered, and other factors. Once the entire year has been completed, the year itself can be summarized in the relevant state variables.

In some cases it is unnecessary to analyze every month of the year. Depending on the kind of person in question, it is sometimes possible to determine directly what his year's activities will be, without examining his month by month behavior. Certain aspects of a person's state vector will determine to what extent the detailed analysis can be bypassed.

The first of these determinants is age. If a person is younger than 14 it is simply assumed that he is a student in primary school. His education status remains unchanged at "less than grade 9", and it is assumed that he spends 40 weeks in school and the remaining 12 in the non-labour force. By definition he has no "April Activity Status" since he is too young to be considered a participant.

When a person reaches age 14 he is treated in a more complex manner. It is assumed in this case that he enters grade 9 in the first month of the simulated year (April). He then goes through the month by month simulation to be discussed below. It might be objected that the person should start grade 9 in September, and in fact this is what



the month by month simulation will finally show. He is initially placed there in April, however, simply for reasons of modelling convenience. This question will be discussed further below.

If a person is older than 14 then the way in which his year is analyzed depends on the employment category he is in. As mentioned earlier, the employment category is a state variable designed to keep track of a person's relation to the labor force. It indicates, for example, whether a person has retired, when he retired, whether he is in pensionable employment, and whether he will be subject to unemployment. It is thus necessary to examine this variable before beginning the monthly labor force simulation.

If a person has retired in the year previous to the one being simulated, it is necessary to change his "TYPE" so as to indicate that he is now in his second year of retirement. Thus type's 40 and 50 are changed to 4 and 5. This change is necessitated by the fact that in his first year of retirement a person must be assigned an initial private pension. In the second and subsequent years of his retirement this person's pension is assumed to be unchanged. It is therefore necessary to distinguish the first year of retirement from subsequent years of retirement.

If a person is retired then the relevant state variables can be assigned directly. It is simply assumed that he spends 52 weeks in the non-labor force and that he will remain in the non-labor force in all subsequent years as well.



If a person's "TYPE" is given as 140, 150, 240, or 250 then the year being simulated is his second year as a labor force participant, and it is necessary to divide these by 10 for the remaining years of the simulation. The reason for this is similar to that necessitating the difference between the first and second year of retirement. The fact that his "Type" is a factor of 10 too high indicates that last year was his first year as a full time labor force participant. At that time it was necessary to know this fact, because a first year participant had to be assigned an initial income in the Market Income Block. Henceforward the distinction is no longer necessary.

It is next necessary to determine whether or not a person will retire in the year being simulated. We distinguish two kinds of individuals. The first, whose "TYPE" ends in a five, is a person who is eligible for a private pension on retirement. This kind of person is eligible for retirement between the ages of 60 and 65 inclusive. Whether or not he will retire in the year being simulated is determined by sampling from a probability distribution. The second kind of individual is one who is not eligible for a private pension. All of these persons are assumed to retire at age 65. For those persons who have never been in the labor force, and are hence "TYPE 3's", it is of course unnecessary to test for retirement. If their age is 65 however, they are designated as "TPYE 4", since it is no longer necessary to test month by month whether they will enter the labor force.





The final means by which a person may bypass the monthly simulation depends on whether he is a Class A individual (TYPE is 14 or 15) or a Class B individual (TYPE is 24 or 25). A Class A individual is assumed always to be employed. He cannot leave the unemployment category until he is retired. He is thus assumed to spend 52 weeks in the employment state, his April activity remains "employed", and his education status remains unchanged.

All of the above cases are those in which the person does not pass through the month to month simulation for one reason or another. The majority of people, however, are in employment categories 24, 25, or 3 (full time labor force participants subject to unemployment, and non-full time labor force participants 14 years and over), and their behavior must be analyzed on a monthly basis. The purpose of the monthly simulation is to determine the person's activity in every month of the year being simulated as well as April of the subsequent year, and to update his education status if he enters school in September. It is first necessary, however, to make a change in the wage variable of these persons. Since the person may not be fully employed, it is necessary, in determining his annual wage, to know both the number of weeks in which he is employed and his weekly wage rate. The Activity Block determines the former, and the Market Income Block determines the latter. Since the Market Income Block proceeds on the basis of weekly wage rate transitions, it is important that an individual's last year's weekly wage not be lost. It is thus calculated in the Activity Block (by dividing WAGES by WKEMP) and stored in



the WAGES location until it can be updated in the income block. The Activity Block now calculates a new WKEMP which is then used with the new weekly wage rate calculated in the Income Block to calculate a new annual wage.

The month to month simulation now proceeds as follows. If the person is 14 then he is assumed to be in grade 9 for every month of the year with the exception of the summer months. The effect of this will be to show the person as spending April through June and September through March in "school", and will leave him in grade 9 in April of the subsequent year. He will then be a year older and will in all likelihood graduate to grade 10 the following September. Strictly speaking, the simulation treats him as being in grade 9 during three months in which he is actually in grade 8 (April through June of the year in which he is 14), but this is irrelevant insofar as the "yearly" state variables are concerned. The "weeks in school" etc. are still tabulated correctly.

For all persons older than 14 the simulation considers transitions among three states: "employment", "unemployment", and "generalized non-labor force (GLF)". The person is in one of these three states during the input month, and he moves to one of these three states for the output month, according to a transition probability which depends on his age, sex, marital status, and region. This kind of transition will occur to most people and is made possible by defining all school states to be within the GLF. A person simply passes from one month to the next until the whole year is completed.



For people who are in school, however, certain exceptions and assumptions relating to the above procedure must be noted. First, if a person is in school it is assumed he cannot drop out. Thus from September through March, transitions from GLF (which in this case is equivalent to some school state) to employment or unemployment are not allowed. But for input months April through August normal transitions take place, thus allowing students to move into the labor force during the summer months.

For transitions in the two months May and June, it is necessary to distinguish between NLF and school for those who enter the GLF state in these months. It will be recalled that GLF includes both of the former states. The required distinction is made as follows. It is assumed that GLF means NLF except for those persons who were in school in the input state. For these latter individuals GLF in the output month is assumed to mean the same school state that the person was in during the input month. That is, a transition from GLF to GLF for students is recognized as being a transition, say, from Grade 11 to Grade 11. For those who are not students it is a transition from NLF to NLF.

We now come to the question of placing a person in school. It is first assumed that the only people for whom it is possible to start a new school year are those who pass into the GLF state in September. The problem then reduces to distinguishing between those who enter some school state and those who enter NLF. This is done on the



basis of a probability that is conditional on the person's age, sex, marital status and April Activity status. If it turns out that a person does enter a school state, a transition matrix determines what his new school state will be, and his education status is updated. The transition matrix is described below in 5.5.1.

Once the monthly simulation has been completed it is a simple matter to total up the number of weeks spent in each of the four states: employment, unemployment, non-labor force, and school. It remains then to assign "TYPE" to those who become new full-time labor force participants in the year just simulated. A person is assumed to have joined the labor force if his present "TYPE" is 3, and if in the present year he spent zero weeks in school and some weeks either employed or unemployed. Type is assigned on the basis of education, province, and sex. Class A persons (whose TYPE will be either 140 or 150) are distinguished by education. These are the people who will never be unemployed and Class A status is assumed to apply to all male university graduates. Whether or not a person is a participant in a private pension plan (thus distinguishing between the "40's" and the "50's") is determined on the basis of a sex-province distribution.

### 5.3 The Labor Force Model

#### 5.3.1 The Basic Model

The Activity Block centers around a Markov-chain model of the labor force which is heavily dependent on a





similar labour force model developed by Donald Dawson and Frank Denton at McMaster University\*. The idea behind the Denton-Dawson model is to abstract from the complex of factors that describe both sides of the market (the labor - leisure choice, wage rates, job vacancies, aggregate demand, etc.) by assuming that all of these factors are adequately represented by the unemployment rate. Thus both supply and demand in the labor market are considered only implicitly. The problem that the model sets out to solve is that of describing, month by month, the labor force activities of certain kinds of persons, given only an unemployment rate that applies to all persons.

To be more specific, assume that at any given time a person must be in one of three mutually exclusive states. He must be either employed (E), unemployed (U), or in the non-labor force (N). Assume further that a person can only move from one of these states to another at the beginning of a month, and that he must then "reside" in that state for the whole month. Assume finally that we wish to know what a person does through some time period, say a year, and that we know in which of the 3 states he is in at the beginning of the period. Then if we knew the probability of moving from one state to another in each of the given months, we could trace out a month by month history of the individual for the year. Calculating these probabilities is the purpose of the Denton-Dawson Model.

---

\* F.T. Denton and D.A. Dawson, "The OTA Simulation System", Report prepared for the Department of Manpower and Immigration, June 1971.



For any given pair of months, 9 probabilities need to be calculated, and these can be designated by a transition matrix:

$$\begin{array}{rcc}
 & & \text{State in month } t+1 \\
 & & \begin{array}{ccc} E & U & N \end{array} \\
 \begin{array}{l} \text{state} \\ \text{in} \\ \text{month } t \end{array} & \begin{array}{l} E \\ U \\ N \end{array} & \begin{bmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{bmatrix}
 \end{array}$$

Thus  $p_{ij}$  is the probability of moving from state  $i$  in month  $t$  to state  $j$  in month  $t+1$ . This probability is assumed by Denton-Dawson to depend on two sets of factors. The first set is the demographic characteristics of the individual, specifically his age and sex. Thus if the population is broken down into 9 age groups and 2 sexes, 18 of the above matrices would have to be calculated for each pair of months being considered. Age and sex we denote as stratification variables.

The second set of factors can be called functional variables. Given a certain stratification of the population, what will the required probabilities depend on? Denton-Dawson assume they will be functions of three things: the average level of unemployment in the two months being considered, the change in unemployment in the two months, and a seasonal factor indicative of the months themselves.

For each age-sex group then, the following equation can be specified:



$$p_{ijt} = L_0 + \sum_{K=1}^{11} L_K M_K + L_{12} \frac{(U_t + U_{t+1})}{2} + L_{13} (U_{t+1} - U_t)$$

where  $p_{ijt}$  is the transition probability between state  $i$  in month  $t$  and state  $j$  in month  $t+1$ .  $M_K$  ( $K = 1, 2, \dots, 11$ ) is a dummy variable that has value 1 if the calendar month is  $K$  ( $K$  is defined to be 1 for January, 2 for February, etc.) and is 0 otherwise. And  $U_t$  is the unadjusted Canadian unemployment rate in month  $t$ .\*

The equation includes among its explanatory variables the mean and difference of unemployment rates in two consecutive months. This is precisely equivalent to using the two unemployment rates themselves ( $U_t$  and  $U_{t+1}$ ) rather than linear combinations of them. The given specification is more useful, however, from the point of view of interpreting labor force behavior. And since the mean and difference of unemployment rates are uncorrelated, the problem of multicollinearity that would otherwise arise is avoided.

Denton-Dawson estimated these equations by making use of data\*\* for the 96-month period starting with December 1961 - January 1962, and ending with November-December 1969. The nine age groups he used were 14, 15-16, 17-19, 20-24, 25-34, 35-44, 45-54, 55-64, and 65-69. With a separate time series for each age group, each sex, and each of the 9 probabilities, he had 162 equations to estimate in all. The equations thus estimated were found to provide a good statistical model of the labor force (in terms of  $R^2$ ,  $t$ -tests,  $F$ -tests, etc.). Denton-Dawson also found that test simulations for

---

\* It will be noted that the equation includes only eleven seasonal dummy variables. This is because it is not possible to estimate a regression equation that has both a constant term as well as a set of dummy terms that sum exactly to unity. i.e. The least squares normal equations must not be dependent. The decision as to which month to drop is quite arbitrary. Exactly the same values for the dependent variable will be generated regardless of which dummy variable is dropped.

\*\* These data were taken from the Labor Force Survey.



periods of up to 10 years, in conjunction with actual historical unemployment rates of the 1960's, reproduced quite closely the actual patterns of movement of the Canadian E, U, and NLF series, as recorded by the monthly DBS Labor Force Survey.

In dealing with transition probabilities, two difficulties are frequently encountered. The first is that since they are probabilities, they must lie in the unit interval. That is,  $0 \leq p_{ij} \leq 1$  for all  $i$  and  $j$ . Secondly, since the three states are both mutually exclusive and exhaustive, it is necessary that each row of the transition matrix sum to one. That is,  $\sum_{j=1}^3 p_{ij} = 1$  for all  $i$ . This latter constraint is satisfied by the original time series data, and it is also satisfied by the generated probabilities because of a convenient property of least squares estimation: if a set of dependent variables is subject to a linear equality constraint and if these variables are all regressed on an identical set of independent variables, any estimates of the dependent variables derived from the regression equations will also be subject to the linear constraint. But the unit interval constraint need not necessarily be satisfied. It is possible to generate negative probabilities, and probabilities that are greater than one. Any cases of this happening are handled by the computer program, which sets the aberrant probability to either zero or one and then normalizes the relevant row.

Estimation of the required transition matrices requires exogenous input of seasonally unadjusted unemployment rates. This is quite inconvenient, since one is far more likely to





have a "feel" for seasonally adjusted rates. Or one would like to input a "low" set of rates and then compare the results of these with the results generated by "medium" or "high" rates. Again, seasonally adjusted rates would be far more convenient for this purpose.

Denton-Dawson solve this difficulty by generating a set of equations that will translate an adjusted rate into an unadjusted rate. The equation is,

$$U_t = \delta_{0K} + \delta_{1K} U_t^* + \delta_{2K} t + \delta_{3K} t^2 + \delta_{4K} t^3$$

where  $U_t$  is the unadjusted unemployment rate,  $U_t^*$  is the seasonally adjusted rate, and  $t$  represents time. There are 12 such equations, one for each month  $K$ . The cubic trend polynomial is included to allow for autonomous shifts in seasonal patterns over time. These equations were estimated from time series data in the 1959-69 period, and were found to perform very well. It is thus possible to specify seasonally adjusted unemployment rates as input to the model.

Although the Denton-Dawson model just described was found to work well in simulating age-sex distributions of employment experience over the historical period in which the equations were estimated, it is in some respects not entirely suitable for the purposes of POLSIM. Three difficulties arise. The first is that the model makes no allowance for structural changes in the labor market over time. If, for example, the efforts of government programs toward correcting seasonal jumps in unemployment figures have been increasingly successful over the last 10 years, then Denton-Dawson's equations will



result in an upward bias in the winter unemployment totals when we come to simulate the future. Second, to allow for a more realistic simulation of wage income, POLSIM distinguishes between two kinds of persons. That is, Class A persons who are assumed to never become unemployed, and whose wage income transitions alone thus determine what their annual income will be; and Class B persons who are subject to unemployment, and hence whose weeks of employment and weekly wage rate transitions combined will determine their annual wage income. The problem this creates in the context of the Denton-Dawson model is that their probabilities apply to all persons in the labor force. POLSIM assumes that the probabilities will only apply to a subset of the population (the Class B persons), and it is therefore necessary to adjust the generated probabilities to account for this. The third difficulty is that the Denton-Dawson model only stratifies the probabilities on age and sex. This is a serious limitation since a person's employment experience depends on other variables as well, most notably region and marital status. If one is interested in the most accurate micro-simulation possible, it becomes necessary to disaggregate the probabilities on these variables as well.

In the following sections we discuss the methods which were adopted to overcome each of the three difficulties mentioned above: the problem of adjusting the model to account for structural shifts in the labor market, the problem of modifying the Denton-Dawson model to account for the fact that POLSIM distinguishes different classes of persons, and the problem of disaggregating the Denton-Dawson model.



### 5.3.2 Model Adjustment to Account for Labor Market Structural Shift

The problem here can be stated quite simply. One generates simulation equations based on data from some historical period. When some part of this period is then simulated, or when some future period is simulated, the simulated figures generated do not match exactly the corresponding actual figures. There are of course many reasons for this. But one very important reason may be that the world has in some way "changed" since the period in which the parameters were estimated, and therefore the historically derived parameters are not altogether suitable. It is reasonable to believe that this is the case with the labor market. Government stabilization programs, changing educational levels of the labor force, increased participation rates on the part of women, increased bureaucratization, increased job security engendered by unions, and the increased attractiveness of unemployment made possible by changes in such programs as unemployment insurance, all lead to a different underlying structure of the labor market. The problem is to account for this in the model.

There are two basic approaches that can be made to this adjustment problem. The first is to adjust the regression coefficients so that they will accurately generate the historical probabilities that correspond to a period as close to the one being simulated as possible. The second is to adjust the coefficients so that they generate probabilities that in turn generate employment-unemployment-non labor force distributions that correspond to those observed in the most recent period. The former method was the one adopted, because it was the most convenient and straightforward. The



latter method is conceptually much more difficult, but could provide a better result. A preliminary analysis of how one might approach the second method is included in Appendix D. It could be incorporated in future versions of the model. The method employed in the current version of POLSIM is given below.

The data Denton-Dawson used to estimate their probabilities ended with the months November-December 1969. Data does exist, however, for the full year from December '70-January '71 to November-December of 1971. An easy check on the regression equations, then, is to generate simulated probabilities for the year 1971 and compare these with the actual probabilities for that year. This was done, using measured 1971 unemployment rates, and significant differences were observed. The adjustment was then quite simple.

Let  $p_i$  = the observed probability in month  $i$

$q_i$  = simulated probability in month  $i$

and  $p_i - q_i = D_i$

All that was necessary was to adjust the regression equations so that they generated  $p$  rather than  $q$ . Since there would be a different  $D_i$  for every month, and since there is a dummy variable in each month, the dummy variables were all increased by the relevant difference. That is,

if  $r_i$  is the original dummy variable coefficient in month  $i$ ,  
and  $s_i$  is the new dummy variable coefficient in month  $i$ ,  
then  $s_i = r_i + D_i$ .

This adjustment was carried out for every month and for all age-sex groups.





- 117 -

Two sets of regression coefficients thus existed, and they were compared by using each set to simulate the period from April 1972 to April 1973. The adjusted equations performed much better as can be seen by referring to the validation section that follows.

### 5.3.3 Adjustment for Class of Person

The Denton-Dawson probabilities apply to the whole population. POLSIM, however, distinguishes two kinds of individuals. "Class A" individuals are assumed to be such that they never enter the unemployment state. They will always be either in NLF or employment. "Class B" individuals will be subject to unemployment and will hence be the only ones who make normal transitions among the three labor force states. Class B individuals must thus absorb all of the unemployment states, while Class A individuals will absorb a large proportion of the employment states. Since the transition matrices will now apply to Class B persons only, and since we still wish to generate the same totals in each state that we would generate if they applied to the whole population, it is necessary to adjust the Denton-Dawson probabilities upwards and downwards, as the case may be. We proceed as follows:



1. Assume the initial population is given as follows:

$E(t)$  = total number of employed individuals at time  $t$ .

$E_1(t)$  = number of Class A individuals employed at time  $t$ .

$E_2(t)$  = number of Class B individuals employed at time  $t$ .

$U(t)$  = number of individuals unemployed at time  $t$ .

$N(t)$  = number of Class B individuals unemployed at time  $t$ .

$N(t)$  = number of individuals in NLF at time  $t$ .

$N_1(t)$  = number of Class A individuals in NLF at time  $t$ .

$N_2(t)$  = number of Class B individuals in NLF at time  $t$ .

2. Let  $p_{ij}$  = probability of moving from state  $i$  to state  $j$  as calculated by Denton - Dawson

where  $i, j, = e, u, \text{ or } n$ .

3. Let  $p_{ij}^!$  = adjusted probability of moving from state  $i$  to state  $j$

where  $i, j = e_1, e_2, u, n_1, n_2$

4. A Priori, we can state that the adjusted transition matrix we are interested in will contain some zeros. They are designated in the following matrix:

$$\begin{array}{c}
 t + 1 \\
 \begin{array}{c} e_1 \quad e_2 \quad u \quad n_1 \quad n_2 \\
 \begin{array}{c} e_1 \\ e_2 \\ u \\ n_1 \\ n_2 \end{array} \left[ \begin{array}{ccccc}
 ? & 0 & 0 & ? & 0 \\
 0 & ? & ? & 0 & ? \\
 0 & ? & ? & 0 & ? \\
 ? & 0 & 0 & ? & 0 \\
 0 & ? & ? & 0 & ?
 \end{array} \right]
 \end{array}
 \end{array}$$



There are thus 13 probabilities to be calculated:  $p'_{e_1 e_1}$   $p'_{e_1 n_1}$   
 $p'_{e_2 e_2}$   $p'_{e_2 u}$   $p'_{e_2 n_2}$   $p'_{ue_2}$   $p'_{uu}$   $p'_{un_2}$   $p'_{n_1 e_1}$   $p'_{n_1 n_1}$   $p'_{n_2 e_2}$   $p'_{n_2 u}$   
 $p'_{n_2 n_2}$

5. To derive these unknown probabilities from the existing Denton-Dawson probabilities, it may be noted that the total number of people moving from one Denton-Dawson state (E, U, N) to another Denton-Dawson state (E, U, or N) must be equal to the same number of people moving between the same two states in the new model.

Thus: (a) Employment - Employment

total number moving from E to E in Denton-Dawson  $p_{ee}^E(t)$

total number moving from E to E in POLSIM is

$$p'_{e_1 e_1} E_1(t) + p'_{e_2 e_2} E_2(t)$$

$$\therefore p_{ee}^E(t) = p'_{e_1 e_1} E_1(t) + p'_{e_2 e_2} E_2(t) \quad (1)$$

(b) Employment - Unemployment

$$p_{eu}^E(t) = 0 \cdot E_1(t) + p'_{e_2 u} \cdot E_2(t)$$

(2)

$$\Rightarrow p'_{e_2 u} = \frac{p_{eu}^E(t)}{E_2(t)}$$

(c) Employment - Non-Labor Force

$$p_{en}^E(t) = p'_{e_1 n_1} E_1(t) + p'_{e_2 n_2} E_2(t) \quad (3)$$



(d) Unemployment - Employment

$$p_{ue} U(t) = p'_{ue_2} U(t)$$

$$\Rightarrow p'_{ue_2} = p_{ue} \quad (4)$$

(e) Unemployment - Unemployment

$$p_{uu} U(t) = p'_{uu} U(t)$$

$$\Rightarrow p'_{uu} = p_{uu} \quad (5)$$

(f) Unemployment - Non-Labor Force

$$p_{un} U(t) = p'_{un_2} U(t)$$

$$\Rightarrow p'_{un_2} = p_{un} \quad (6)$$

(g) Non-Labor Force - Employment

$$p_{ne} N(t) = p'_{n_1 e_1} N_1(t) + p'_{n_2 e_2} N_2(t) \quad (7)$$

(h) Non-Labor Force - Unemployment

$$p_{nu} N(t) = p'_{n_2 u} N_2(t)$$

$$\Rightarrow p'_{n_2 u} = \frac{p_{nu} N(t)}{N_2(t)} \quad (8)$$

(i) Non-Labor Force - Non-Labor Force

$$p_{nn} N(t) = p'_{n_1 n_1} N_1(t) + p'_{n_2 n_2} N_2(t) \quad (9)$$

We also can see from examining the transition matrix that

$$p'_{e_1 e_1} = 1 - p'_{e_1 n_1} \quad (10)$$

and 
$$p'_{n_1 n_1} = 1 - p'_{n_1 e_1} \quad (11)$$





6. We thus have 11 equations in 13 unknowns. By substituting equation (10) into equation (1) and equation (11) into equation (9) we eliminate the unknowns  $p'_{e_1 e_1}$  and  $p'_{n_1 n_1}$ . We can then write the remaining 9 equations in 3 independent sets.

Set 1

$$p'_{ee} E(t) = (1 - p'_{e_1 n_1}) E_1(t) + p'_{e_2 e_2} E_2(t) \quad (1')$$

$$p'_{en} E(t) = p'_{e_1 n_1} E_1(t) + p'_{e_2 n_2} E_2(t) \quad (2')$$

Set 2

$$p'_{e_2 u} = \frac{p'_{eu} E(t)}{E_2(t)} \quad (3')$$

$$p'_{ue_2} = p_{ue} \quad (4')$$

$$p'_{uu} = p_{uu} \quad (5')$$

$$p'_{un_2} = p_{un} \quad (6')$$

$$p'_{n_2 u} = p_{nu} \frac{N(t)}{N_2(t)} \quad (7')$$

Set 3

$$p'_{ne} N(t) = p'_{n_1 e_1} N_1(t) + p'_{n_2 e_2} N_2(t) \quad (8')$$

$$p'_{nn} N(t) = (1 - p'_{n_1 e_1}) N_1(t) + p'_{n_2 n_2} N_2(t) \quad (9')$$

7. It can be seen that the equations in Set #2 determine five of the unknown probabilities explicitly.



The transition matrix thus looks like:

$$\begin{array}{c}
 \text{Year } t+1 \\
 \begin{array}{ccccc}
 e_1 & e_2 & u & n_2 & n_2
 \end{array} \\
 \text{Year } t \begin{array}{c}
 e_1 \\
 e_2 \\
 u \\
 n_1 \\
 n_2
 \end{array}
 \left[ \begin{array}{ccccc}
 ? & 0 & 0 & 1-p'_{e_1 e_1} & 0 \\
 0 & ? & \frac{p_{eu} E}{E_2} & 0 & ? \\
 0 & p_{ue} & p_{uu} & 0 & p_{un} \\
 ? & 0 & 0 & 1-p'_{n_1 e_1} & 0 \\
 0 & ? & \frac{p_{nu} N}{N_2} & 0 & ?
 \end{array} \right]
 \left. \begin{array}{l} \\ \\ \\ \\ \end{array} \right\} \begin{array}{l} \\ \text{unknowns } p'_{e_1 e_1} p'_{e_2 e_2} p'_{e_2 n_2} \\ \\ \text{unknowns } p'_{n_1 e_1} p'_{n_2 e_2} p'_{n_2 n_2} \\ \end{array}
 \end{array}$$

8. From Set 1 we can determine the 3 unknowns in the top 2 rows. To do this it is necessary to specify from some outside source the probability  $p'_{e_1 e_1}$

9. Similarly Set 3 determines the unknowns in the bottom 2 rows. Here it is necessary to first specify  $p'_{n_1 e_1}$

10. For age classes less than 65 years it is reasonable to specify

$$p'_{e_1 e_1} = 1.0 \Rightarrow p'_{e_1 n_1} = 0.0$$

and for classes more than 65 years

$$p'_{e_1 e_1} = 0.0 \Rightarrow p'_{e_1 n_1} = 1.0$$



11. A not unreasonable hypothesis is that

$$p'_{n_1 e_1} = 0.0 \text{ for all age classes}$$

This in effect means that we allow no transitions for Class A individuals from NLF to employment. In fact, we assume that  $N_1 = 0$  initially, i.e., the number of Class A individuals in NLF initially is zero. This means that "Class A" individuals are defined (initially) to be people in "preferred" occupations only. A person not in the labor force would be just that: we wouldn't distinguish "preferred" NLF from "ordinary" NLF. The only people to ever enter  $N_1$  would then be Class A people who retire. And, they will never return to  $E_1$ . This procedure ensures some consistency between the assumptions here and the assumptions in 11 directly above.

12. If the assumptions discussed in 10 and 11 are made, the following two matrices can be derived:

(a) for ages < 65

$$\begin{array}{c}
 \begin{array}{ccccc}
 & e_1 & e_2 & u & n_1 & n_2 \\
 e_1 & \left[ \begin{array}{ccccc}
 1 & 0 & 0 & 0 & 0 \\
 0 & \frac{p_{ee}^{E-E_1}}{E_2} & \frac{p_{eu}^E}{E_2} & 0 & \frac{p_{en}^E}{E_2} \\
 0 & p_{ue} & p_{uu} & 0 & p_{un} \\
 0 & 0 & 0 & 0 & 0 \\
 0 & p_{ne} & p_{nu} & 0 & p_{nn}
 \end{array} \right]
 \end{array}
 \end{array}$$



(b) for ages > 65

$$\begin{array}{c}
 e_1 \quad e_2 \quad y \quad n_1 \quad n_2 \\
 \begin{array}{l}
 e_1 \\
 e_2 \\
 u \\
 n_1 \\
 n_2
 \end{array}
 \left[ \begin{array}{ccccc}
 0 & 0 & 0 & 1 & 0 \\
 0 & \frac{p_{ee}E}{E_2} & \frac{p_{eu}E}{E_2} & 0 & \frac{p_{en}E-E_1}{E_2} \\
 0 & p_{ue} & p_{uu} & 0 & p_{un} \\
 0 & 0 & 0 & 1 & 0 \\
 0 & p_{ne} & p_{nu} & 0 & p_{nn}
 \end{array} \right]
 \end{array}$$

13. Once the ratios  $\frac{E}{E_2}$  and  $\frac{E_1}{E_2}$  are determined for a given age-sex group, the existing probability matrices can be transformed into the above "adjusted" matrices. (In practice only the first row of the existing 3X3 matrices, which corresponds to the  $e_2$  row in the adjusted matrices, will be altered. The  $e_1$  and  $n_1$  rows in the adjusted matrices are really deterministic, and hence the transitions they imply can be carried out without recourse to random numbers and the probability matrix.)

14. In principle, the adjustment parameters defined above apply only to a given time period. They depend on the total number of Class A and Class B individuals employed at a given moment in time, and these can be expected to change as time progresses. For the present, however, we assume these changes do not significantly alter the relevant parameters ( $E/E_2$  and  $E_1/E_2$ ), and hence the latter will remain constant over the whole period simulated.





#### 5.3.4 The Disaggregation of Transition Matrices

It will be recalled that the transition matrices in the Denton-Dawson model are stratified on age and sex alone. This entails no particular problem, provided that we do not attempt to assess distributional consequences beyond the age-sex level. If we are interested in finer distributions, however, the Denton-Dawson model loses much of its usefulness. Regional implications, for example, are of particular interest in Canada. But the Denton-Dawson model does not distinguish between regions.

It thus becomes important to try to further disaggregate the Denton-Dawson matrices. Broadly speaking, what we would like to do is select any particular age-sex matrix, and from it produce 10 matrices (one for each marital status-region class). These new matrices would hopefully better reflect the particular classes to which they belong, while at the same time not changing the aggregate age-sex distributions they would produce.

In principle, there does not exist any ideal way to handle this problem. It is simply not possible to create detail not adequately represented in the original Labour Force Survey data. Any attempt to disaggregate data to levels finer than those actually measured will at best yield approximations. It is the accuracy of these approximations that either justify or refute the method. Having said this, we proceed.

Suppose that a transition matrix (for a given age and sex) is give by:

$$p = \begin{bmatrix} p_{ee} & p_{eu} & p_{en} \\ p_{ue} & p_{uu} & p_{un} \\ p_{ne} & p_{nu} & p_{nn} \end{bmatrix}$$



This matrix produces distorted distributions with respect to marital status and region that we would like to correct.

We can define a new transition matrix as:

$$p' = \begin{bmatrix} p_{ee} + \Delta p & p_{eu} - \Delta p & p_{en} \\ p_{ue} + \alpha \Delta p & p_{uu} - \alpha \Delta p & p_{un} \\ p_{ne} & p_{nu} & p_{nn} \end{bmatrix}$$

Where  $\Delta p$  is unknown,  $\alpha$  is some constant (say .1), and the  $p_{ij}$ 's are the same as in the original matrix.

That is, we specify that our new matrix is to be such that the first element in the first row is increased (or decreased) by a constant amount, and the first element in the second row is to be increased (or decreased) by some fraction of this amount. The second element in each row is then to be adjusted so as to make the sum of the first two elements the same as it was originally.

Let us now assume that  $p'$  is a matrix that produces exact results for a given marital status and region class. That is, for a given age-sex class, it is being postulated that a "good" matrix for one of the marital status-region classes can be obtained by slightly adjusting the original matrix in the way outlined above. Similarly, a "good" matrix for another of the marital status-region classes can be obtained by a similar adjustment, with a different parameter (a different  $\Delta p$ ).

The problem is now as follows. Select a particular age-sex group, and hence a particular matrix. Select also one of the ten marital status-region groups. For the above



assumption to be true (the assumption that the new matrix will give exact results) there must exist a  $\Delta p(t)$  such that for the given marital status and region the actual number of employed in period  $t$  will equal the number of employed simulated by the new matrix in period  $t$ , and the actual number of unemployed in period  $t$  will equal the number of unemployed simulated by the new matrix in period  $t$ . The problem is simply to find  $\Delta p(t)$  for the particular marital status, region, and month chosen. If we repeat the process for every marital status and region, we will generate 10 new matrices (2 sexes x 5 regions) for every one of the old ones. And hopefully these will produce accurate simulations on all four variables: age, sex, marital status, and region.

The derivation of  $\Delta p(t)$  now follows.

1. Let  $E_A(t)$  = number of actual employed persons at time  $t$ .  
(for a given marital status-region group).

$E_N(t)$  = new total simulated at time  $t$  (from new matrix)

$E_S(t)$  = old total simulated at time  $t$  (from original matrix)

$U_A(t)$  = number of actual unemployed persons at time  $t$   
(for a given marital status-region group)

$U_N(t)$  = new total simulated at time  $t$

$U_S(t)$  = old total simulated at time  $t$

The problem is to find  $\Delta p(t)$  such that  $E_A(t) = E_N(t)$  and  $U_A(t) = U_N(t)$  for all  $t$ .

2. We can write:

$$(1) \quad E_N(t+1) = (p_{ee} + \Delta p(t))E_N(t) + (p_{ue} + \alpha \Delta p(t))U_N(t) + p_{ne}N(t)$$



for  $t = 1$ ,

$E_N(1) = E_M(1) =$  measured employed on initial year tape

$U_N(1) = U_M(1) =$  measured unemployed initial year tape

Equation (1) then becomes

$$E_N(2) = p_{ee}E_M(1) + p_{ue}U_N(1) + p_{ne}N(1) + \Delta p(1)(E_M(1) + \alpha U_M(1))$$

$$\text{or } E_A(2) = E_S(2) + \Delta p(1)(E_M(1) + \alpha U_M(1))$$

$$\therefore \boxed{\Delta p(1) = \frac{E_A(2) - E_S(2)}{E_M(1) + \alpha U_M(1)}}$$

3. For  $t = 2, 3, \dots$  we can write equation (1) as:

$$E_N(t+1) = (p_{ee} + \Delta p)E_N(t) + (p_{ue} + \alpha \Delta p)U_N(t) + p_{ne}N(t)$$

$$\text{Setting } E_N(t) = E_A(t)$$

$$U_N(t) = U_A(t)$$

$$E_N(t+1) = E_A(t+1)$$

we have

$$E_A(t+1) = p_{ee}E_A(t) + p_{ue}U_A(t) + p_{ne}N(t) + \Delta pE_A(t) + \alpha \Delta pU_A(t)$$

We can write  $E_A(t)$  as  $E_S(t) + E_A(t) - E_S(t)$

and  $U_A(t) = U_S(t) + U_A(t) - U_S(t)$ , which then gives:

$$\begin{aligned} E_A(t+1) &= p_{ee}E_S(t) + p_{ue}U_S(t) + p_{ne}N(t) + (p_{ee} + \Delta p)(E_A(t) - E_S(t)) \\ &\quad + (p_{ue} + \alpha \Delta p)(U_A(t) - U_S(t)) + \Delta pE_S(t) + \alpha \Delta pU_S(t) \\ &= E_S(t+1) + p_{ee}(E_A(t) - E_S(t)) + p_{ue}(U_A(t) - U_S(t)) \\ &\quad + \Delta p(E_A(t) + \alpha U_A(t)) \end{aligned}$$

Now because we force the total actual labor force to equal the total simulated labor force for purposes of validation (thus ignoring any errors caused by leaks to the non labor force) it is true that





$$E_A(t) + U_A(t) = E_S(t) + U_S(t)$$

$$\text{or } E_A(t) - E_S(t) = U_S(t) - U_A(t) = \beta$$

$$\text{Then, } E_A(t+1) = E_S(t+1) + (p_{ee}-p_{ue})\beta + \Delta p(E_A(t) + \alpha U_A(t))$$

$$\text{or } \Delta p(t) = \frac{E_A(t+1) - E_S(t+1) - (p_{ee}-p_{ue})\beta}{E_A(t) + \alpha U_A(t)}$$

4. It remains to explain the meaning of this equation. First of all, we can note that there are two reasons why the simulated totals differ from the actual totals. The first is simply that the transition matrices, the  $p_{ij}$ 's, are incorrect. The second reason is that as the simulation proceeds, the transition matrices are multiplying state totals that are incorrect. They operate on the totals produced by the incorrect  $p_{ij}$ 's, rather than the actual totals.

In explaining the derived equation it will simplify matters if we assume  $E_A$  is larger than  $E_S$ . That is, we are simulating too few employed people and too many unemployed. Considering the numerator, the term  $E_A(t+1)-E_S(t+1)$  is simply the total deficiency of employed persons being simulated. Because too few employed persons were simulated for period  $t$  (the deficiency being equal to  $\beta$ ),  $p_{ee}\beta$  will be the deficiency in period  $t+1$  arising for this reason from the employment transition. And similarly,  $p_{ue}\beta$  will be the excess



employeds simulated in period  $t+1$  because too many unemployed were simulated in period  $t$ . The term  $(p_{ee}-p_{ue})\beta$  is thus the total deficiency of employeds in  $t+1$  arising from the creation of too many employeds in period  $t$ . The numerator is therefore the deficiency of employed persons in period  $t+1$  that can be explained by the errors in the  $p_{ij}$ 's alone.

Intuitively, one could more easily understand the equation for  $\Delta p(t)$  if it were given by

$$\Delta p(t) = \frac{E_A(t+1) - E_S(t+1)}{E_A(t)}$$

It would simply state that  $p_{ee}$  for period  $t$  to  $t+1$  should be increased by the relative amount that the actual employed in period  $t+1$  exceeds the simulated employed in period  $t+1$ . But this would not be wholly accurate. This is because it would be putting the full burden of adjustment on the  $p_{ee}$  term. And we wish to let the  $p_{ue}$  term absorb part of the adjustment. This explains the 2nd term in the denominator of the derived equation. The more that the second row can account for the deficiency of employeds (the higher are  $\alpha$  and  $U_A(t)$ ), the less the adjustment ( $\Delta p$ ) that falls on the first row.

The addition of this term to the denominator would still not make the intuitive equation correct however. We do not wish to adjust for the full error of actual persons over simulated persons, because we know that part of this error is a cumulative effect caused by previous errors. Thus it is necessary to subtract off this cumulative error, namely the  $(p_{ee}-p_{ue})\beta$  term.



## 5.4 Validation of the Activity Block

### 5.4.1 Theoretical Aspects

The validation of the Activity Block proceeds as a slight generalization of the principles elucidated in section 4.3.1. We will begin by mathematically describing the labor force Markov-process.

We define the following variables:

- (a)  $X_t = \begin{bmatrix} E_t & U_t & N_t \end{bmatrix}$  is the vector of labor force states at time  $t$ . ( $t = 0$  is the initial period)

That is,  $E_t$  = total number of employed people at time  $t$ .

$U_t$  = total unemployed at time  $t$ .

$N_t$  = total non labor force at time  $t$ .

$$(b) \quad P_t = \begin{bmatrix} P_{ee} & P_{eu} & P_{en} \\ P_{ue} & P_{uu} & P_{un} \\ P_{ne} & P_{nu} & P_{nn} \end{bmatrix}$$

= The transition matrix for period  $t-1$  to  $t$ .

$$(c) \quad S_t = P_1 P_2 \dots P_t$$

= the transition matrix from period zero to period  $t$ .

$$\text{Then, } X_1 = X_0 P_1 = X_0 S_1$$

$$X_2 = X_1 P_2 = X_0 P_1 P_2 = X_0 S_2$$

$$X_t = X_{t-1} P_t = X_0 S_t$$



Now if both  $X_0$  and the  $S_t$ 's were known exactly, and if the simulation process was perfect, then the simulated  $X_t$ 's would also be perfect. But none of these conditions hold. Rather, the simulated  $X_t$ 's can be written as

$$\begin{aligned}\hat{X}_t &= (S_t + R_t) (X_0 + \Delta X_0) + \varepsilon_t \\ &= S_t X_0 + R_t X_0 + S_t \Delta X_0 + R_t \Delta X_0 + \varepsilon_t \\ &= X_t + R_t X_0 + S_t \Delta X_0 + R_t \Delta X_0 + \varepsilon_t\end{aligned}$$

Where,

$R_t$  = the error transition matrix from period zero to period  $t$ . This is also called "parameter error"

$\Delta X_0$  = the error in the initial labor force vector, or "initial state error"

$\varepsilon_t$  = the simulation error. This error arises because we sample from a probability distribution using a random number generator. It can be thought of as the same kind of error that arises when we toss a fair coin. We know that if we toss a coin 100 times, we should expect 50 heads and 50 tails. In fact, however, we know that if we performed the experiment many times, we would find that the observed number of heads was binomially distributed, with a mean of 50 and a variance of 25. We would hardly ever hit 50 right on, and the difference above or below 50 could be called the "simulation error".





In the expression for  $\hat{X}_t$  we can ignore the second order term,  $R_t \Delta X_0$ , and write,

$$\begin{aligned}\hat{X}_t - X_t &= \text{Total error} = R_t X_0 + S_t \Delta X_0 + \epsilon_t \\ &= \text{parameter error} + \text{initial state error} \\ &\quad + \text{simulation error.}\end{aligned}$$

Ideally, we would like to distinguish these three kinds of errors in any analysis of simulation results. Unfortunately, this would be both difficult and costly. Since the errors that do result in the Activity Block are quite small (see below), a slightly different approach is therefore taken.

The simulated results are compared with the actual totals in two ways. The first examines the total error. Simulated and actual values for both employment and unemployment are compared, and the percentage error calculated. These comparisons are shown in the top rows of the tables in Appendix D. The total error thus presented is not really indicative of how well the simulation performs, however, since a large proportion of the error is the result of initial state error. In an attempt to correct for this, an "adjusted" comparison is given in the second row of each of the tables. In this comparison, the simulated values are adjusted pro rata so as to add up in total to the actual measured values. That is, the simulated employed and unemployed totals are adjusted either up or down so that their sum will be the same as the sum of the corresponding measured totals.



The adjusted simulation results calculated in this way represent a correction for the initial state error. The model population that POLSIM works with is somewhat smaller than the total Canadian population\*, and comparisons between actual and simulated values are obviously going to reflect this. The adjustment process just brings both populations to the same total size, without changing the distributions within these populations. This does not correct the whole of the initial state error however. If the distribution of the initial population is also in error, in addition to the size of the population, then some initial state error will remain. Inspection of the initial distributions indicates that this latter kind of error does exist, but that it is small. Since correction would involve some rather tenuous assumptions, the effects of this error are ignored in the present analysis.

The difference between the adjusted simulated totals and the actual totals are thus for the most part a result of the combination of simulation error and parameter error. Since the sum of these errors are for the most part very small, no attempt was made to distinguish them.

#### 5.4.2 Validation Results

The Activity Status Block was validated by simulating the thirteen months from April 1972 to April 1973. Simulated labor force aggregates (total unemployed, total employed,

---

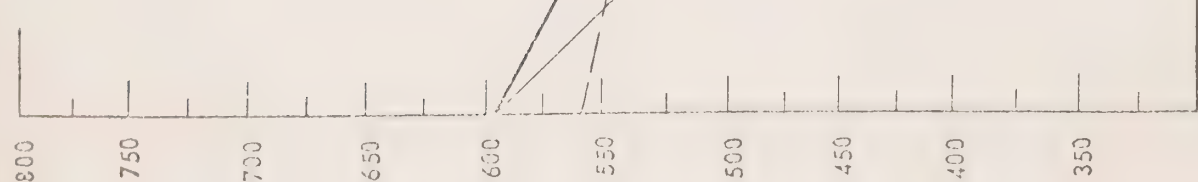
\* Recall from Chapter 2 that the Survey of Consumer Finance does not include persons in institutions, Indians on reservations, or people who live in the Yukon or North West Territories.



# NUMBER OF UNEMPLOYED PERSONS GRAND TOTAL

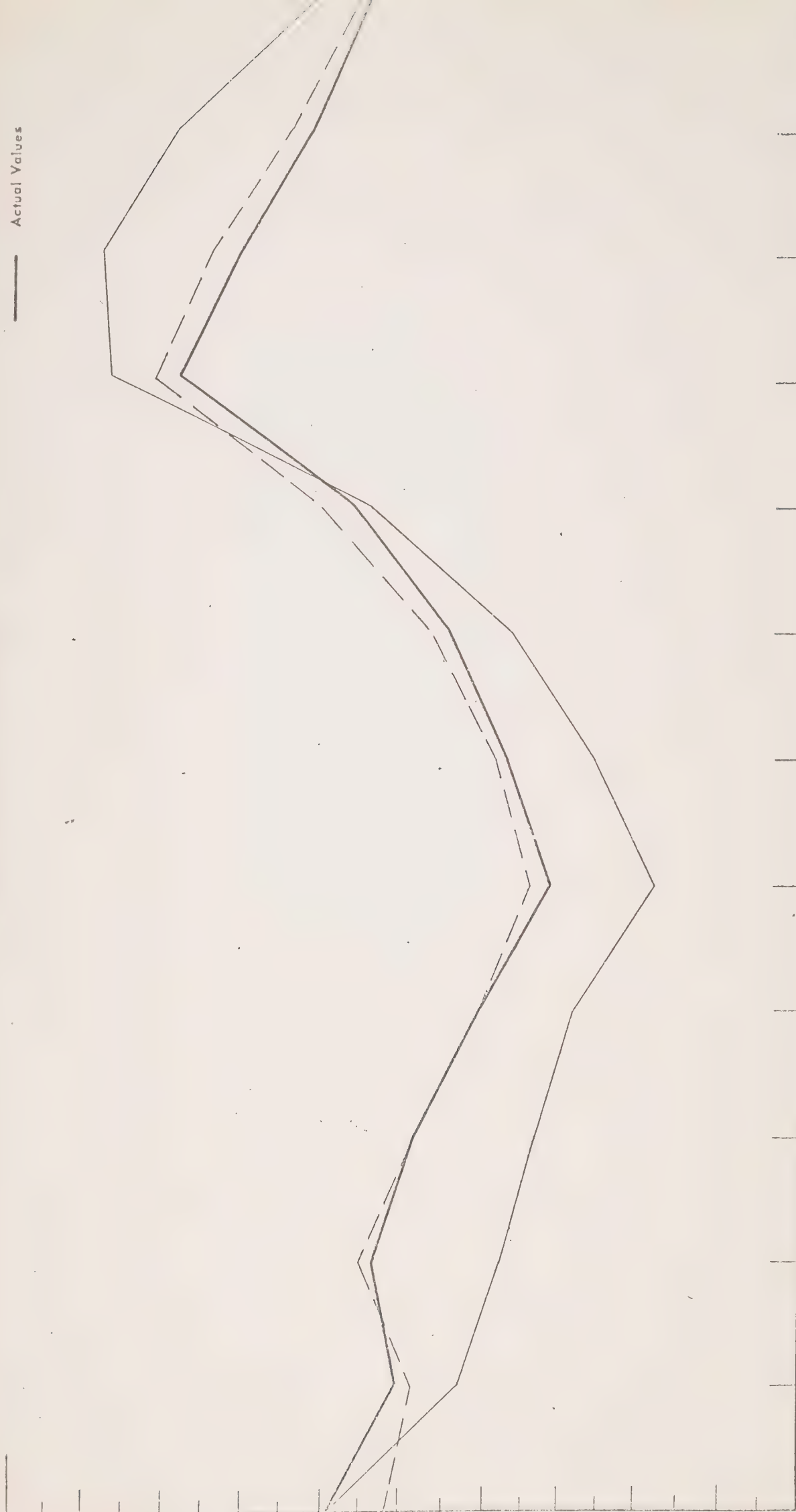
Original Simulation  
Collocated Class Adjustment  
with Disaggregation  
(Final Simulation)  
Actual Values

THOUSANDS



APR  
1972

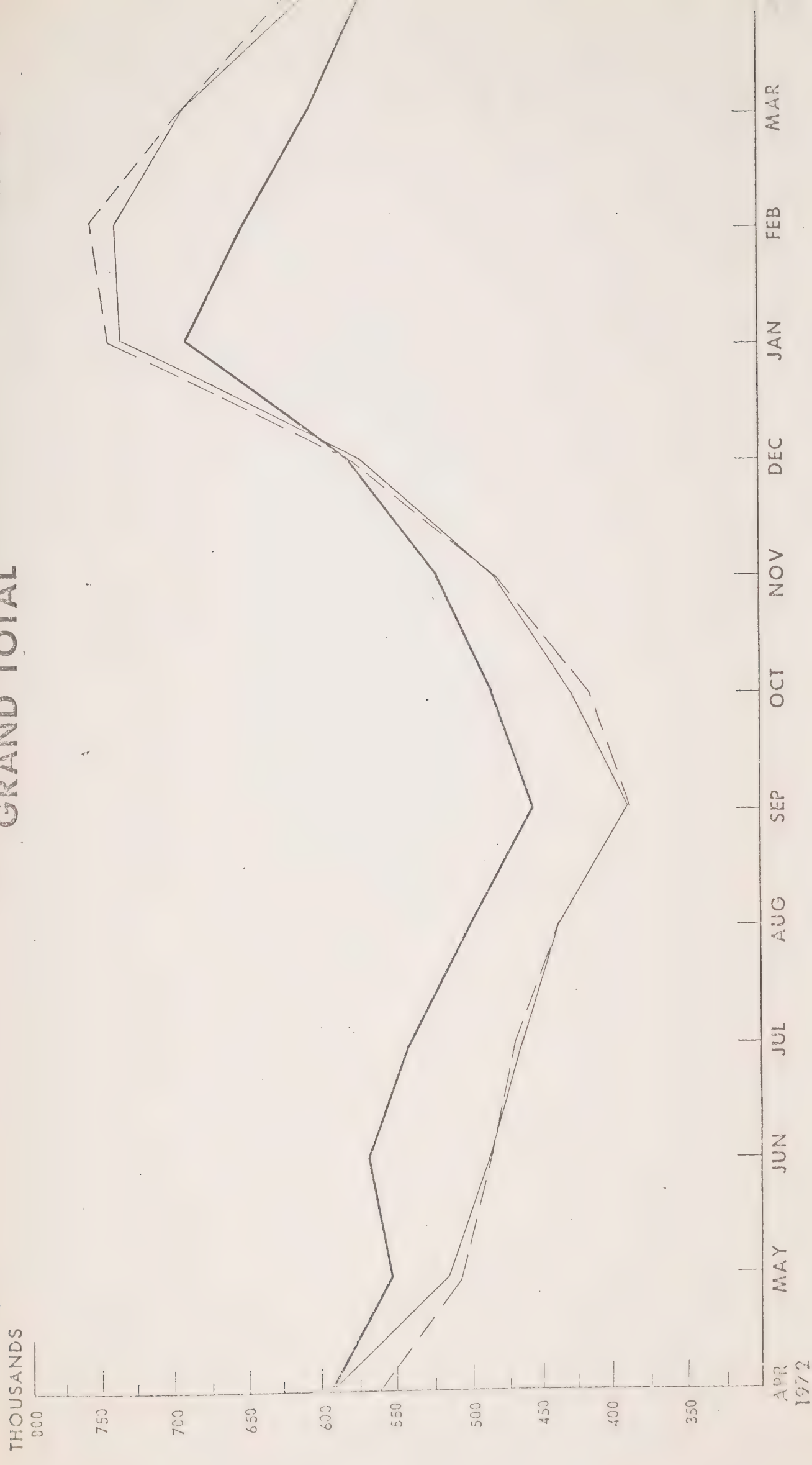
MAY JUN JUL AUG SEP OCT NOV DEC JAN FEB MAR





NUMBER OF UNEMPLOYED PERSONS  
GRAND TOTAL

Original Simulation  
Class A adjustment  
Actual Values







NUMBER OF UNEMPLOYED PERSONS

GRAND TOTAL

Class A adjustment

Calibrat f Class A  
adjustment

Actual Values

THOUSANDS

800

750

700

650

600

550

500

450

400

350

APR

1972

MAY

JUN

JUL

AUG

SEP

OCT

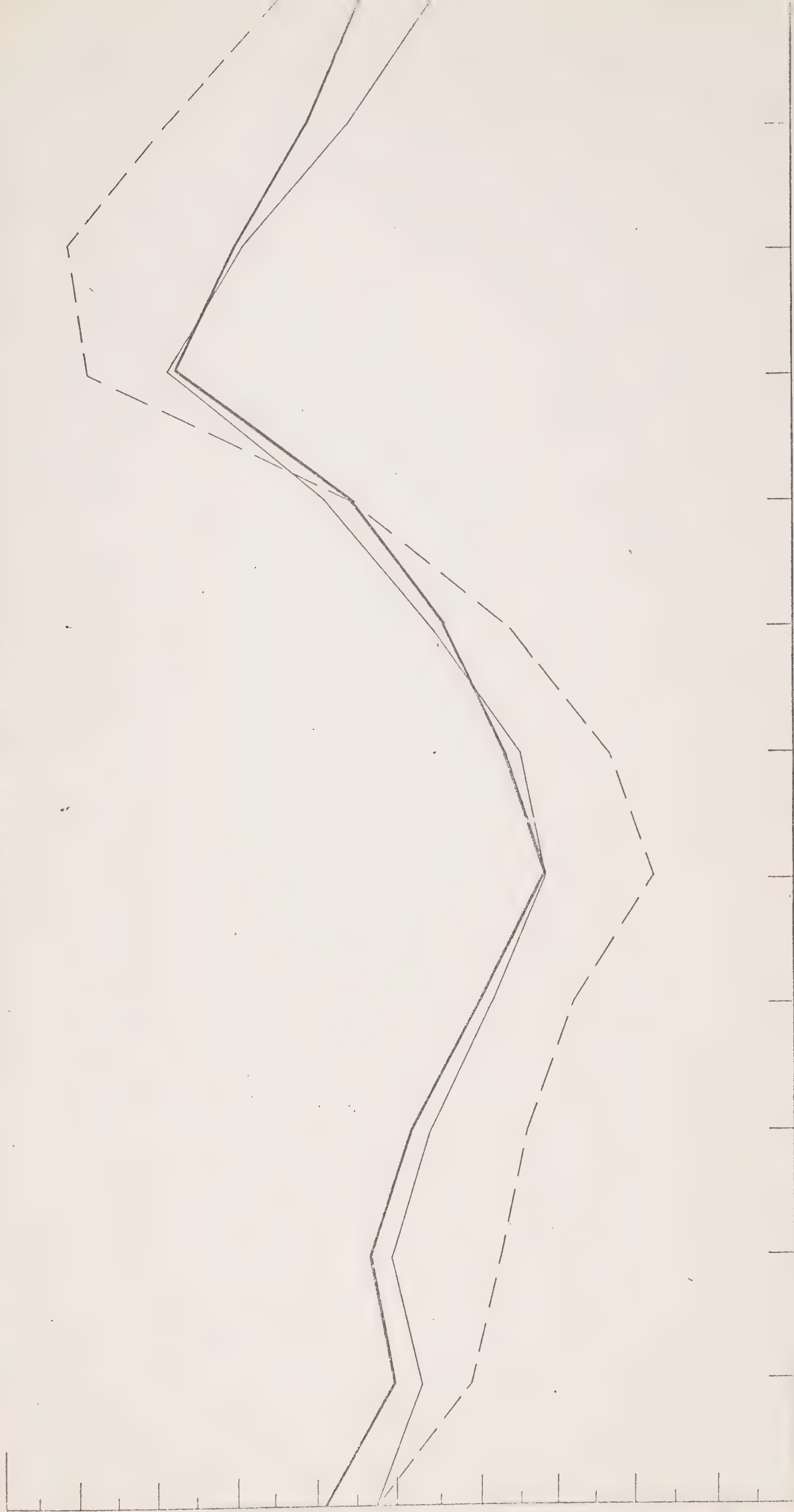
NOV

DEC

JAN

FEB

MAR





NUMBER OF UNEMPLOYED PERSONS  
GRAND TOTAL

Calibrated Class A adjustment  
Calibrated Class A adjustment  
with Disaggregation  
(Final Simulation)  
Actual Values

THOUSANDS

800

750

700

650

600

550

500

450

400

350

APR  
1972

MAY

JUN

JUL

AUG

SEP

OCT

NOV

DEC

JAN

FEB

MAR

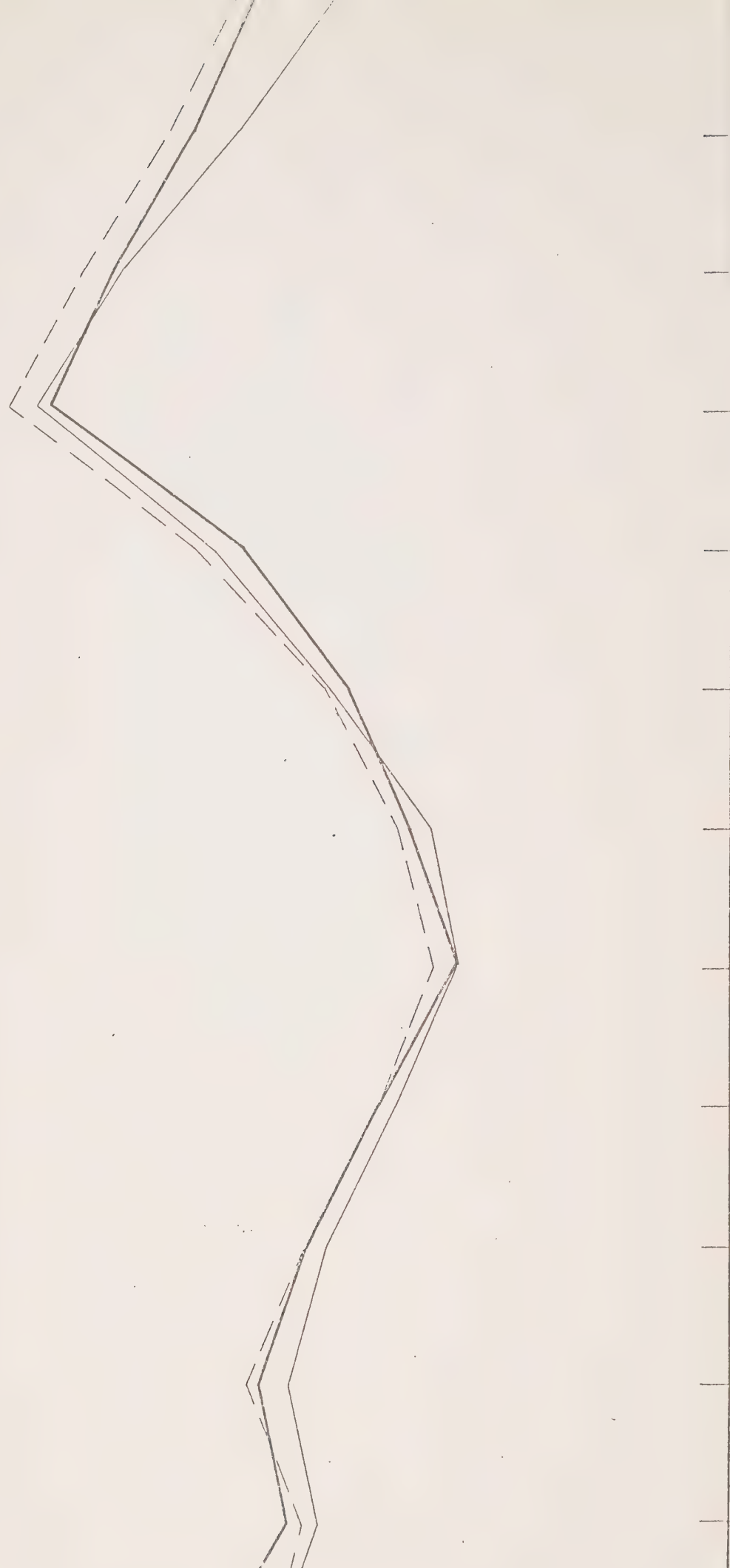
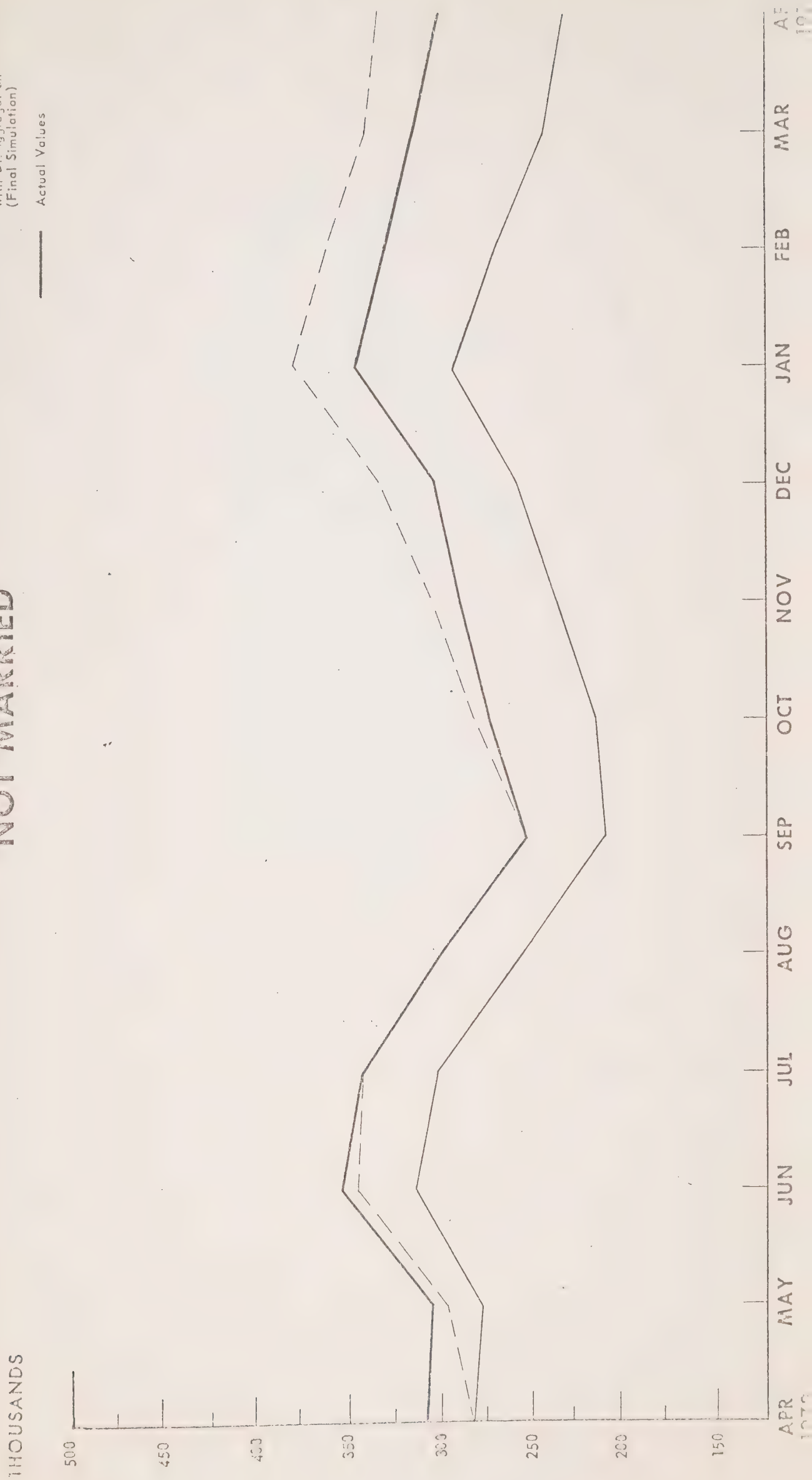




FIGURE 5.3e

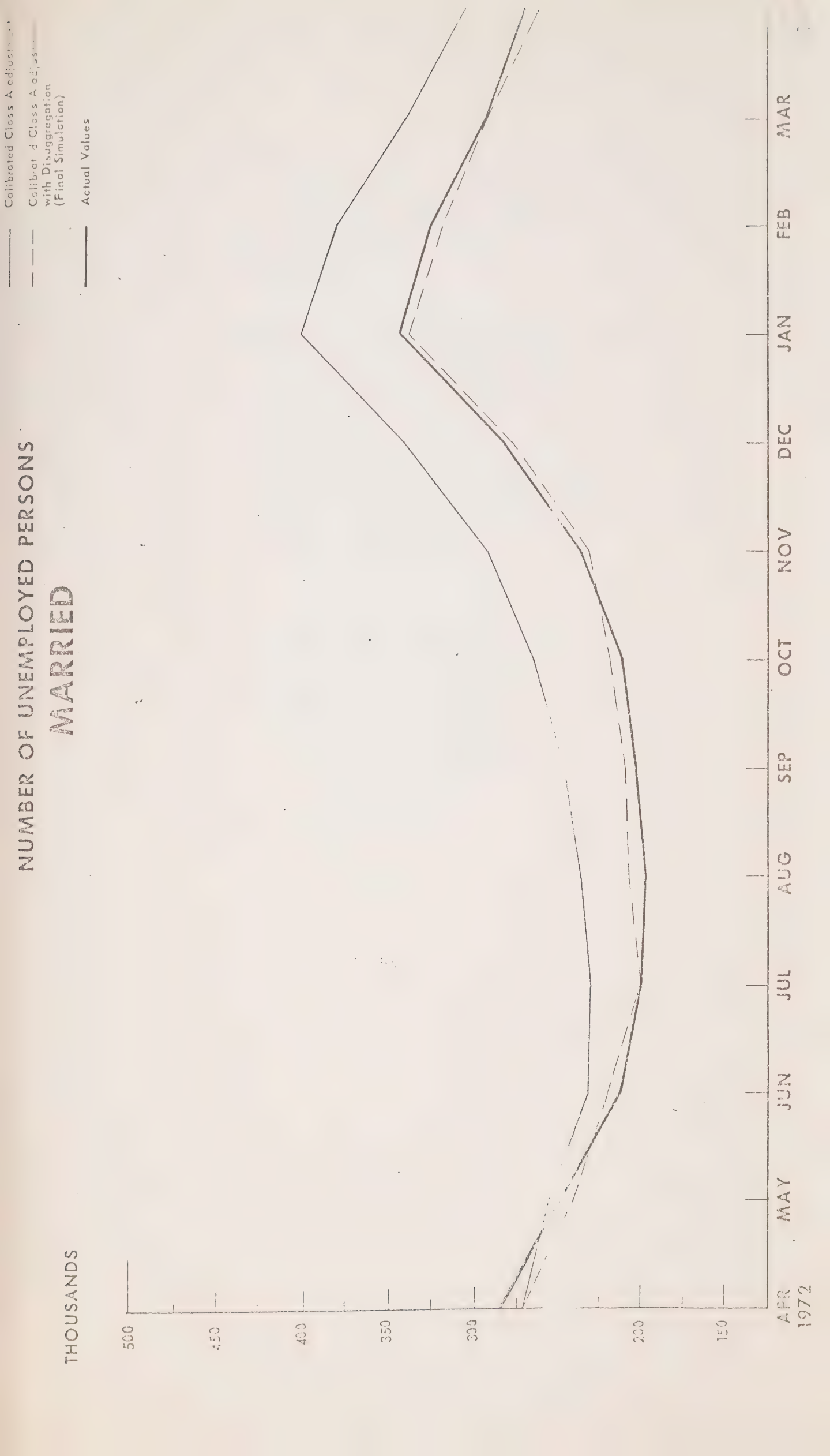
# NUMBER OF UNEMPLOYED PERSONS NOT MARRIED

Calibrated Class A adjustment  
Calibrated Class A adjustment  
with Disaggregation  
(Final Simulation)  
Actual Values





# NUMBER OF UNEMPLOYED PERSONS MARRIED







total male unemployed, etc.) were computed and compared with the corresponding actual values as measured by the labor force survey. The results of this comparison are presented in detail in Appendix D.

The validation can be summarized with reference to the graphs in figure 5.3. The graphs illustrate the difference between the results obtained using the original Denton-Dawson parameters, and the results obtained when the various adjustments to these parameters (see sections 5.3.2 - 5.3.4) were introduced. The graphs also show how the simulated totals compare with the actual totals.

The first four graphs plot the total number of unemployed persons, and show the cumulative and net effect of the various adjustments that were made to the parameters. The first graph compares the original simulation to the final simulation. The final set of parameters, it will be recalled, contains all three adjustments: the adjustment necessitated by the distinction of Class A and Class B persons (the "Class A adjustment"), the adjustment to account for structural shifts in the labor market (the "Calibrated adjustment"), and the adjustment necessary to disaggregate the parameters to region and marital status (the "Disaggregation adjustment"). It can be seen that there is a considerable improvement in the results obtained using the final set of parameters as compared with the original set. The average error with the original parameters was 9.4%. With the adjusted parameters, the average error was 2.3%.



The second graph compares the original simulation with the Class A adjustment. In the absence of simulation error, the two sets of parameters should yield identical results, and in fact they generally do. The only significant difference between the two simulations occurs in the base month. This is not a reflection on the adjustment process itself, which clearly "works" more than adequately. The difference in the initial month can rather be attributed to the fact that the process that distinguishes Class A individuals in the base year population is itself imperfect. Some of those whom we designate as Class A, and hence as always employed, are in fact unemployed in the base month. This is not a serious problem, however, because as can be seen from the graph, the Markov-chain process soon corrects for this small initial error.

The third graph illustrates the effect of calibration. A significant improvement can be seen to have occurred in every month. The average error is reduced from 10.4% to 2.4%.

The last three graphs illustrate the disaggregation adjustment. Over all, as can be seen from figure 5.3d, there is little difference between the results obtained using the calibrated parameters, and the results obtained using the disaggregated parameters. At the highest level of aggregation, the two sets yield virtually the same results. And this is as we would expect. But if we examine the totals for the two marital status classes, we find that the calibrated parameters significantly overestimate the number of unemployed



married people, and significantly underestimate the number of unemployed people. This is again as we would expect, since the calibrated parameters do not distinguish between married and single people. The disaggregated parameters, on the other hand, do make this distinction. And as can be seen from the graphs, they effect a considerable improvement at this lower level of aggregation.

Similar comparisons for different age groups, regions, sexes, and the two marital status classes are given in Appendix D.

## 5.5 Data and Inputs

The activity block uses four different sets of data. The first set relates to the problem of moving people through school. The second set is used to move people through the three labor states (employment, unemployment, and non-labor force) and to determine retirement and participation in pension plans. The third set consists of parameters that are used to make various adjustments in the calculated labor force transition probabilities. And the last set consists of unemployment rates and other parameters that are exogenous input to the labor force transition probability calculations. All of the input data, with the exception of the unemployment rates, is listed in Appendix D.



### 5.5.1 School Transitions

$$1. \quad \underline{PSCL(I,J,K,L) = PSCL(10,2,2,22)}$$

This is the conditional probability that a person in age group I, marital status state J and sex K will be in school in September, given that he was in state L in the previous April and that he will be in either school or the non-labor force in September.

The indices are as follows:

- (a)    1 = 1 if age is 14                      6 if age is 19  
          2                      15                      7                      20-24  
          3                      16                      8                      25-29  
          4                      17                      9                      30-34  
          5                      18                      10                     35-39
- (b)    J = 1 if married  
              2 if not married
- (c)    K = 1 if male  
              2 if female
- (d)    L = 1 Grade 9                      12 Univ 4  
              2 Grade 10                      13 Univ 5  
              3 Grade 11                      14 Univ 6  
              4 Grade 12                      15 Univ 7  
              5 Grade 13                      16 Univ 8  
              6 CAAT 1                      17 Univ 9  
              7 CAAT 2                      18 Univ 10  
              8 CAAT 3                      19 Retraining  
              9 Univ 1                      20 Employed  
              10 Univ 2                      21 Unemployed  
              11 Univ 3                      22 NLF

#### Source

This data is derived from sets of transition matrices compiled from various sources by Leroy Stone of Statistics Canada\*.

---

\* Leroy O. Stone, "Preparation of Some Demographic and Socio-Economic Data Inputs", Statistics Canada Internal Report.





$$2. \quad \underline{FGG(I,J,K,L,M) = FGG(10,2,2,18,18)}$$

These are a set of 40 cumulative transition matrices (18 x 18) that determine how a person moves from one school state to another. There is one matrix for each age (I)-marital status(J)-sex(K) group. L is the input state and M is the output state. The indices are as described under PSCL above. (The input and output states are the 18 school states listed under index L above.)

#### Source

These matrices were also derived from the data compiled by Leroy Stone.

#### 5.5.2 Activity Transitions

The data in this set consists of regression coefficients from which Labor Force transition matrices are calculated. The regression equations were derived by Frank Denton and D.A. Dawson of McMaster University. They have discussed their methodology in several papers\*.

##### 1. The Raw Data

The raw data consists of month-to-month transition matrices disaggregated by age and sex for the years 1959-69. The transition probabilities are for movements among the three

- 
- \* 1. D.A. Dawson, "Report on Data Sources Relevant to Simulation Models of the Canadian Adult Training System".
2. F.T. Denton, "A Simulation Model for Month-to-Month Labour Force Movement in Canada", McMaster University Department of Economics: Working Paper No. 72-11.
3. F.T. Denton and D.A. Dawson, "Some Models for Simulating Canadian Manpower Flows and Related Systems", McMaster University Department of Economics: Working Paper No. 72-14.



states: employment, unemployment, and non-labor force. These data were derived by Denton and Dawson from the labor force survey and were used by them to derive the regression coefficients which are the basic input to the Activity Block Model.

$$2. \quad \underline{C(I,J) = C(12,5)}$$

These are the five regression coefficients used to calculate the seasonally unadjusted unemployment rate in month I from the adjusted rate for that month. The regression equation is the cubic trend polynomial:

$$\begin{aligned} \text{VRATE}(I) = & ((C(I,1)*\text{URATE}(I)/100) + C(I,2)*5 + \\ & C(I,3)*25 + C(I,4)*125. + C(I,5))*100 \end{aligned}$$

where  $\text{URATE}(I)$  = seasonally adjusted rate in month  
I (as %)

$\text{VRATE}(I)$  = unadjusted rate in month I (as %)

and I = 1 for Jan.

= 2 for Feb.

etc.

#### Source

These coefficients were derived by Denton and Dawson\*.

---

\* F.T. Denton and D.A. Dawson, "The OTA Simulation System", Report prepared for the Department of Manpower and Immigration, June 1971.



$$3. \quad \underline{A(I,J,K,L,M) = A(3,3,9,2,14)}$$

These are the regression coefficients for calculating the transition matrices. The regression equation is:

$$P(I,J,K,L,N) = A(I,J,K,L,1) + A(I,J,K,L,M+1) + \\ A(I,J,K,L,13)*UBAR + A(I,J,K,L,14)*DELU$$

Where

I = row of the transition matrix

J = column of the transition matrix

K = age of the person

1	14	4	35-44
2	15-16	7	45-54
3	17-19	8	55-64
4	20-24	9	65-69
5	25-34		

L = sex

1	male
2	female

M = Calendar month from which simulation occurs,

e.g. M = 1 means transition from Jan. to Feb.

M = 11 means transition from Nov. to Dec.

UBAR = average unadjusted unemployment rate

$$= (VRATE(R) + VRATE(R+1))/2$$

where R = month from which simulation occurs.

DELU = change in unadjusted unemployment rate

$$= VRATE(R+1) - VRATE(R)$$

#### Source

These coefficients, which were derived by Denton and Dawson\*, are given in Appendix D.

---

\* Ibid.



#### 4. The Calibrated A Matrix

The calibrated A matrix is exactly equivalent to the A matrix described in (3) above. The only difference is that the dummy values (values 2 through 11) have been adjusted so as to force the regression equations to yield exactly the observed probabilities for the year 1971. These new coefficients are also listed in Appendix D.

#### 5. Private Pension Eligibility

The following raw data exists from which private pension eligibility may be inferred:

a.	<u>Pension Plan Members</u>		<u>Total Paid Workers</u>
	(000's)		(000's)
	<u>Male</u>	<u>Female</u>	
Nfld	29.82	8.25	117.
PEI	5.77	2.67	24.
NB	46.9	16.7	220.
NS	70.3	24.9	173.
PQ	549.0	191.5	1871.
ONT	907.4	314.7	2688.
MAN	90.5	31.8	300.
SASK	60.77	24.4	226.
ALTA	125.2	53.0	498.
BC	178.4	63.0	722.

#### b. Total paid workers by sex:

Male = 4483, female = 2356

where a "paid worker" is defined as "those workers employed in a situation where an employer-employee relationship exists". It thus excludes all unpaid family workers and the self-employed.





These data are derived from: Statistics Canada, "Pension Plans in Canada 1970", #74-401 and from the Labor Force Survey #71-001 (March '73).

From this data we wish to derive a province-sex participation rate distribution.

Let  $W(I,M)$  = number of male paid workers in prov.I

$W(I,F)$  = number of female paid workers in prov.I

$WT(I)$  = total paid workers in province I  
 $= W(I,M) + W(I,F)$

Assume  $\frac{W(I,M)}{W(I,F)} = a = \text{constant for all provinces}$

$$= \frac{4483}{2356}$$

$$= 1.9$$

$$\text{Then } W(I,F) = \frac{WT(I)}{1+a} = \frac{WT(I)}{2.9}$$

$$\text{and } W(I,M) = \frac{WT(I)a}{1+a} = \frac{WT(I)}{1.53}$$

Let  $P(I,M)$  = number of male pension plan members in province I

$P(I,F)$  = number of female pension plan members in province I

Then the participation rates are given by:

$$R(I,M) = \frac{P(I,M)(1.53)}{WT(I)}$$

$$R(I,F) = \frac{P(I,F)(2.9)}{WT(I)}$$

Using these relations the participation rates are as follows:

<u>Province (I)</u>	<u>R(I,M)</u>	<u>R(I,F)</u>
NFLD (1)	.39	.20
PEI (2)	.37	.32
NB (3)	.33	.22
NS (4)	.62	.32
PQ (5)	.45	.30
ONT (6)	.52	.34
MAN (7)	.46	.31
SASK (8)	.41	.31
ALTA (9)	.38	.31
BC (10)	.38	.25

These may be compared with Canada wide participation rates:

Total = .41, males = .465, females = .31



6. Retirement Age for those in Receipt  
of Private Pensions

The Statistics Canada publication, "Pension Plans in Canada 1970" (#74-401) supplies data on the "normal retirement age of people who have private pension plans. Statistics Canada defines "normal retirement age" as "the earliest age at which a member may retire as a right and receive immediately his full accrued pension without reduction, although it is not necessarily the age at which he leaves the service of the employer". The data is given in the following table:

<u>Age</u>	<u>Males(%)</u>	<u>Females(%)</u>
60 or less	10.3	28.6
61-64	0.6	1.3
65	76.6	53.9
66 and over	<u>1.9</u>	<u>1.0</u>
	89.4	84.8

The totals do not sum to 100% because some plans provide for optional normal retirement ages based on some combination of age and minimum service retirements. From this data we have to derive actual retirement ages.

It can be assumed, to begin with, that most people in the "optional normal retirement age group" will actually retire somewhere between 60 and 65. Also, most of those whose "normal" retirement age is less than 60, or is between 60 and 64 will undoubtedly actually retire somewhere in the 60-65 age bracket. Thus to determine actual retirement age, two arbitrary decisions must be made: (a) those with optional retirement ages must be distributed in the 60-65 age brackets; and (b) those whose "normal" retirement age is less than 65 must also be distributed in the 60-65 age bracket. That is,



for males, 21.5% (10.3+6+11.6) of the population must be so distributed; and for females, 45.1% (28.6+1.3+15.2) must be distributed.

If we distribute these uniformly, and assume that all those whose "normal" retirement age is 65 or greater actually do retire at 65, then we can derive the following table.

<u>Actual Retirement Age</u>		
<u>Age</u>	<u>Males(%)</u>	<u>Females(%)</u>
60	3.6	7.5
61	3.6	7.5
62	3.6	7.5
63	3.6	7.5
64	3.6	7.5
65	82.0	62.5
	<u>100.0</u>	<u>100.0</u>

It must be noted that this age distribution applies only to those who will be in receipt of private pensions. For the others, who will receive no private pension, it will be assumed that retirement takes place at age 65.

### 5.5.3 Adjustment Parameters

#### 1. CHANGE (I,J,K,L) - CHANGE (3,3,9,2)

This is the calibration term used to adjust the December-January transition probabilities. Since no dummy term exists for the December-January transition (because it would lead to matrix singularity) it is not possible to simply adjust the A matrix for this particular month. The calibration term must be added explicitly.



The indices are:

I = row of matrix for which term applies

J = column of matrix for which term applies

K = age group (see description of A matrix)

L = Sex: 1 - males, 2 - females

Source

This data is derived by calculating the probabilities generated for the December (1970) - January (1971) transition by the original regression coefficients and then comparing these with the actual probabilities as obtained from the Labor Force Survey.

2. RATIO (I,J) - RATIO (2,9)

These are the parameters that adjust the labor force transition probabilities to account for the Type 1 - Type 2 distinction. The indices are:

I = 1 for total employed/type 2 employed

I = 2 for type 1 employed/type 2 employed

J = age group

1	14	6	35-44
2	15-16	7	45-54
3	17-19	8	55-64
4	20-24	9	65+
5	25-34		

The array applies to males only.





Source

This data was derived from the 1971 Survey of Consumer Finance.

3. DADJ (I,J,K) - DADJ (2,5,12)

These are the adjustment parameters used to adjust the age-sex transition matrices so as to account for region and marital status differences. DADJ(I,J,K) is the parameter that accounts for the adjustment for marital status I, region J and month K (transition is from month K to month K+1).

I = 1 not married

2 married

J = 1 Atlantic

2 Quebec

3 Ontario

4 Prairies

5 British Columbia

K = 1 April, .... 12 March

Source

These parameters were derived from data for the year April 1972 through April 1973. The labor force survey in each of those months gave data on the total employed and total unemployed broken down by region and marital status. Simulation for the same months gave slightly different totals. Comparison of the two sets of data gave the adjustment parameters, as discussed in the section 5.3.4.



#### 5.5.4 Variable Parameters

The following parameters are read in from cards. They define the particular year(s) being simulated.

1. IMONTH

This is the first calendar month from which a transition is to occur. Normally this month is April and  $IMONTH = 4$ .

2. NMONTH

The total number of months in the simulation. Normally one full year is simulated, April through April, and  $NMONTH = 13$ . This means that the simulation will contain  $NMONTH - 1$  transitions plus one input month.

3. URATE(13)

This is the seasonally adjusted unemployment rate vector for the 13 months expressed as a percentage. That is, if  $URATE$  for month  $I$  is 6%,  $URATE(I) = 6.0$ .

4. KYR

This is the year being simulated. If the base year is 1971, the first year simulated will be 1972 (April 1972 to April 1973) and  $KYR$  will be input as "1972".



## 6. The Market Income Block

### 6.1 The Market Income Model

#### 6.1.1 Introduction

The POLSIM model divides an individual's total annual income into four main components: employment income, property income, retirement income, and other money income. Employment income is income in the form of wages and salaries, military pay and allowances, and net income from farming, fishing, and other forms of self-employment. Property income is divided into two subcomponents: dividends and other investment income. Dividends are self explanatory. Other investment income consists, generally, of income from fixed face value assets such as bonds, deposits, and savings certificates, and all other forms of investment income. Retirement income consists of private pensions, superannuation, and annuities. Other money income includes income from roomers and boarders, alimony, gifts, and income from any other source not mentioned above.

The purpose of the Market Income Block is to update these income variables, on an annual basis, as the person moves from year to year through the simulation. Broadly speaking, this updating process consists of two general problems. The first is to assign initial component incomes to a person if he is "eligible" and if the particular component being examined was zero in the previous year. (A "component" of income is one of the four sources mentioned above.) The second problem is to determine transitions on each of the income components that were not zero in the previous year. Both of these two kinds of processes are generally handled by either a deterministic function, a time-invariant



stochastic function, or a combination of these. The exact meaning of "eligibility" in the various cases, and the way in which the several transitions are effected, will now be discussed.

#### 6.1.2 Overview of the Model

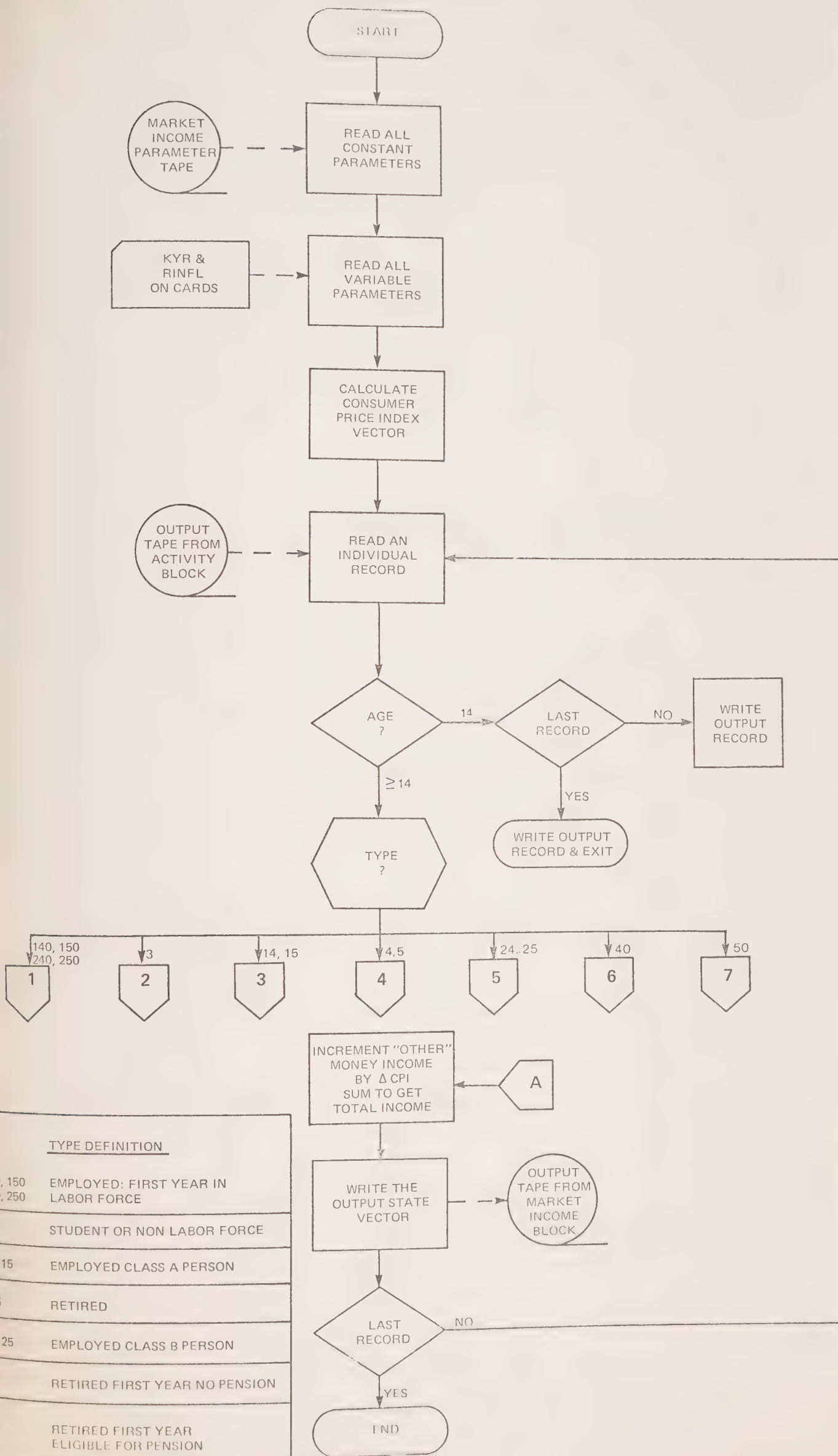
A micro-flow chart of the Market Income Block is shown in figure 6.1. It can be seen that the block begins by reading and assembling all of the parameters that will be necessary to update an individual's income variables. All of the constant parameters, which consist of transition matrices, initial income arrays, growth factors, etc., are read in from a single data tape. They are then simply stored for use by the relevant subroutines. The variable parameters are the particular year being simulated and the rate of inflation that is assumed to apply throughout the simulated period. Both of these variables are read in from cards.

After reading in the rate of inflation, the program is able to calculate a consumer price index vector. This is simply the consumer price index for the 15 years from 1967 to 1981. It will be used to calculate money growth in certain components of a person's income. The CPI vector is the last parameter necessary to effect individual transitions. All of the other sets of data have already been stored, and the program is able to begin to read and process individual records.

Before the income updating processes begin, the program checks to see if the individual is less than 14 years of age. Children less than 14 are assumed by POLSIM to receive no income, and consequently the income variables of all such children are not updated.





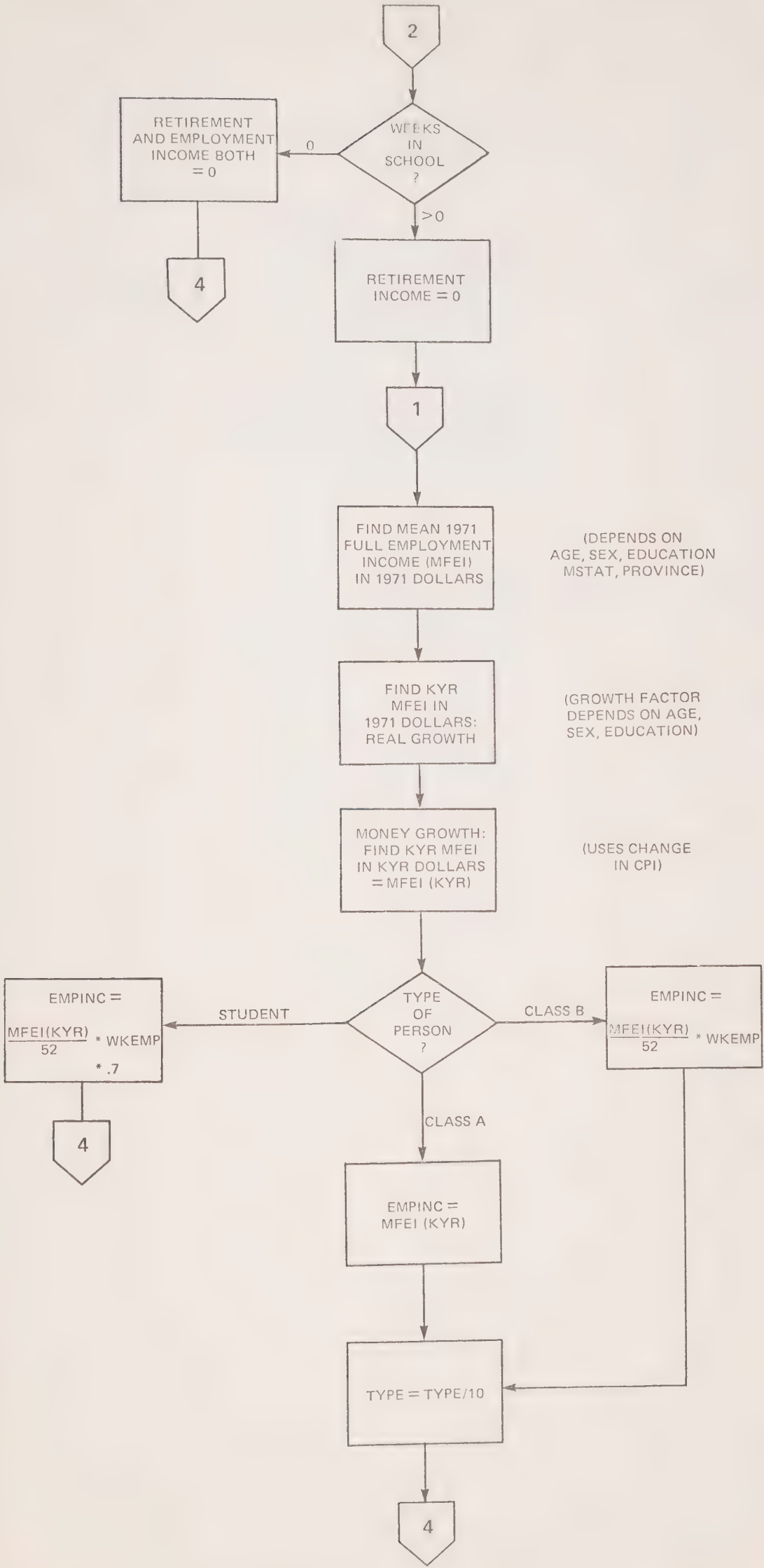


TYPE DEFINITION

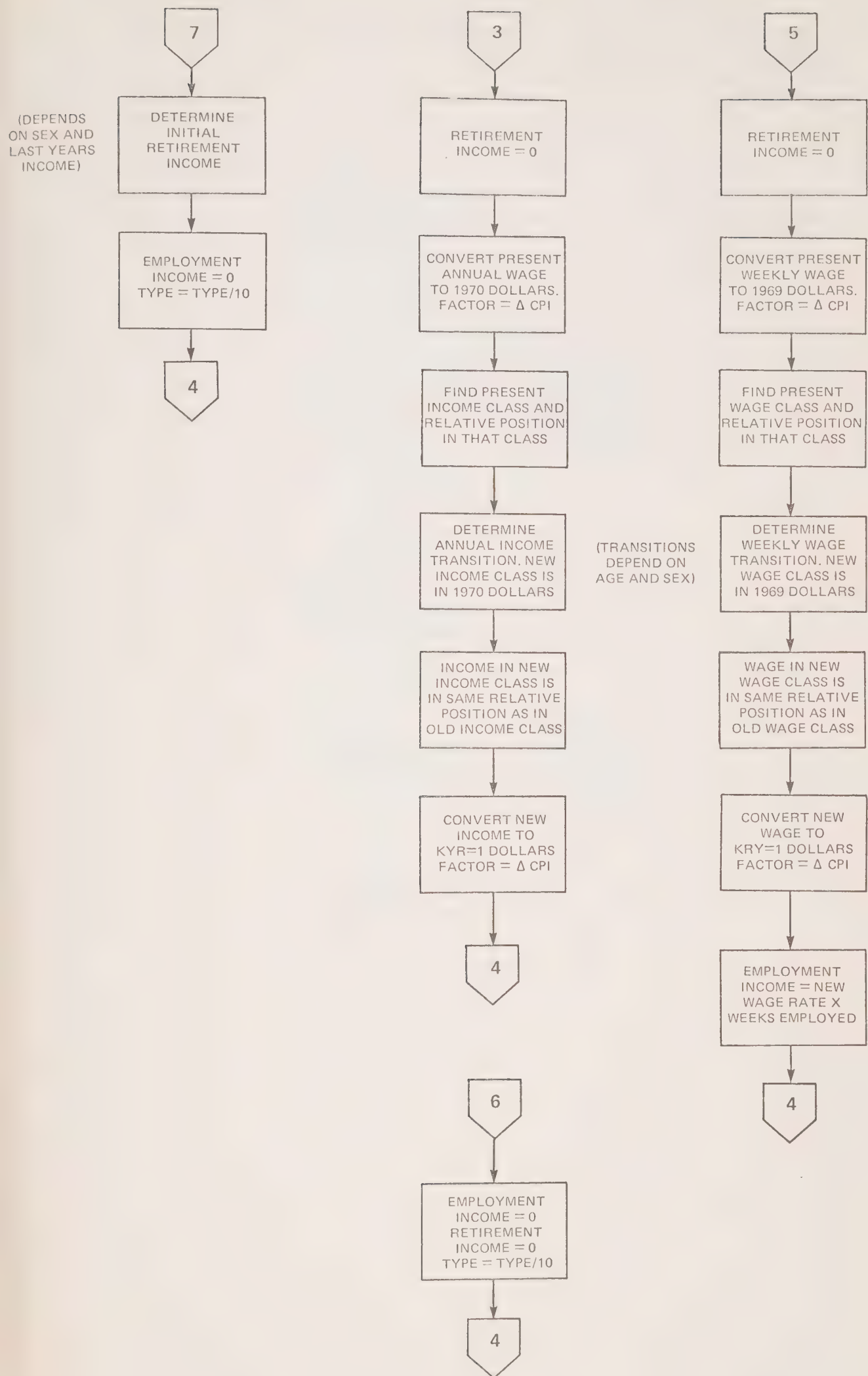
140, 150 240, 250	EMPLOYED: FIRST YEAR IN LABOR FORCE
3	STUDENT OR NON LABOR FORCE
14, 15	EMPLOYED CLASS A PERSON
4, 5	RETIRED
24, 25	EMPLOYED CLASS B PERSON
40	RETIRED FIRST YEAR NO PENSION
50	RETIRED FIRST YEAR ELIGIBLE FOR PENSION



INITIAL AND STUDENT INCOMES

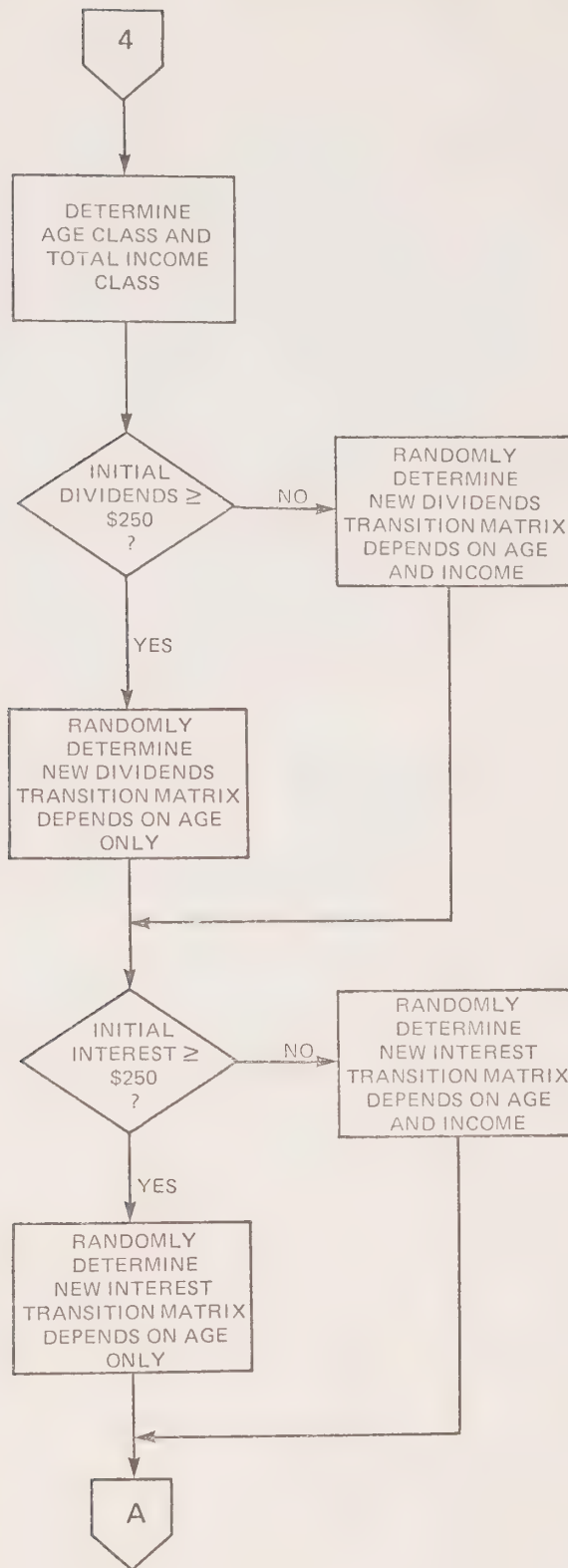








# PROPERTY INCOMES







All other persons are assumed to be eligible for employment, property and other kinds of income. The way in which the transitions on these kinds of income are handled will be discussed in Section 6.3 below. For employment and retirement incomes, which depend on one's relation to the labour force, a variety of processes exist. The exact transformations that these two sources of income are subject to depend on certain elements in the individual state vector. That is, they depend on just what "kind" of person the given individual is.

The main criterion for determining how a person's employment and retirement income will be updated is his "TYPE". If his TYPE is greater than 100 then he is a person who has entered the labour force for the first time in the current year, and he consequently must have an initial income assigned to him. His retirement income will be zero. If his TYPE is 3, then he is either a student or a full-time member of the non-labour force (and not retired). In the former case, an employment income must be assigned if the student worked during the summer. In the latter case, no employment income is assigned. In neither case is a retirement income assigned. If his TYPE is 4 or 5 then he is retired. His retirement income is assumed to be constant, and his employment income is zero. If his TYPE is 14 or 15 then he is a Class A employed person. He is assumed to be employed for the full year, and his employment income is determined by an annual employment income transition. His retirement income is of course zero. If the person is TYPE 24 or 25 then he is a Class B member of the labour force. The Activity block has previously determined how many weeks he has worked, and the Market Income block now determines any change in his weekly wage.



His employment income for the year is then the product of these two variables. His retirement income is zero. If the person is TYPE 40, then this is his first year of retirement (he has reached age 65), and he is not eligible for a private pension. In the present version of the model he is then assumed to have no private retirement\* or employment income at all. If a person is TYPE 50 then he has just retired and he is eligible for a private pension. This pension is calculated and becomes the whole of his retirement income. His employment income is zero.

### 6.1.3 The Income Processes

#### 1. Initial and Student Incomes

This process applies to all persons who must be assigned an initial employment income, and to all students with some summer employment.

A mean full employment income in 1971 dollars is first computed deterministically, on the basis of the person's age, sex, education, marital status, and province. Real economic growth is then applied to this income, depending on an exponential growth factor. This growth factor is a function of the age, sex, and education of the person under consideration. Multiplication by the growth factor results in current year real income (the current year is the year being simulated) expressed in 1971 dollars. This is then inflated by the change in CPI, giving current mean full employment income in current dollars.

---

\*This does not comprehend public programs such as CPP or QPP.



The person's actual employment income is then calculated. If the person is a Class A individual, his employment income remains as the income just calculated. Since Class A persons are assumed to never become unemployed, no changes from the full employment income are necessary. If the person is Class B, then it is possible that he may not have worked a full year. The full employment income is therefore converted to a weekly rate, and the individual's employment income becomes the product of his wage rate and the number of weeks he worked. If the person is a student he is treated in almost the same manner as a Class B person. The only difference is that a multiplicative factor of .7 is introduced into the wage rate to account for the differential between student summer wages and the equivalent wage paid to full-time employees.

## 2. The Annual Income Transition Process

The central feature of this process is an annual wage transition matrix, stratified on age and sex, which defines the probability of a person moving from one income class to another. The income classes of this matrix are defined in 1970 dollars. Since the person's income will be expressed in current year dollars, the first step in the transition process is to deflate the person's income to 1970 dollars. This is done by simply multiplying the person's income by the ratio of the relevant CPI's. It is then necessary to find the particular 1970 income class into which the person falls, and his position in this class, relative to the lower bound of the class. Once this is done, the person's new income class is determined stochastically. A random number is drawn, and the new class is assigned depending on the value of this number and the cumulative transition probability distribution relevant to the particular initial income class in



question. The person's exact income in the new income class is then calculated. He is assigned the same relative position in the new class that he held in the previous year's class. The income thus computed is still expressed in 1970 dollars. It is therefore inflated to current dollars to reflect the change in the CPI over the particular given period.

### 3. The Weekly Wage Transition Process

The process here is virtually identical to the one described above. The only difference is that the transition matrices are defined by weekly wage rates in 1969 dollars.\* The transition process from the old wage rate in current dollars to the old wage rate in 1969 dollars, and thence to the new wage rate in 1969 dollars and the new wage rate in current dollars is exactly the same, mutatis mutandis, as described above. Once the new wage rate is calculated, the person's employment income for the year becomes his weekly wage rate multiplied by the number of weeks he worked.

### 4. The Retirement Income Process

The assignment of initial retirement income is quite straightforward. A probability distribution, contingent on sex, defines the percentage of last year's annual earnings that the person will receive as pension. This distribution is sampled and the relevant percentage determined. The person's retirement income is then just this percentage of his last year's annual earnings.

---

\*The reason why the weekly transition matrices are in 1969 dollars and the annual transition matrices in 1970 dollars is that no data existed on weekly wage transitions for the years 1970-71. The last years of data for weekly transitions are 1969-70 and the last years for annual transitions are 1970-71. (cf. Section 6.2.1)





## 5. The Property Income Processes

It will be recalled that property income consists of two components, dividends and other investment income. The transition processes that apply to dividends are exactly equivalent to those that apply to other investment income. The only difference between the two is that the probabilities defining the two processes are different. But the processes themselves, the definition of the cells in the transition matrices, and the variables that the matrices are stratified on, are exactly the same. The discussion below will therefore be in terms of one process only: one applicable to "property income". As the flow chart in figure 6.1 indicates, however, what really takes place in the program are two sequential identical processes, the first for dividends and the second for other investment income.

The updating of "property income" can be divided into two sections, depending on how much initial property income the person in question has. The first applies only to persons whose property income is less than \$250 and embraces those persons who are moving from zero to non-zero property incomes for the first time. For people in this group, transitions to other classes depend on both age and income. A random number samples a cumulative probability distribution which is contingent on the person's age and income. The sampling defines the new property income class, and hence the new property income (which is taken to be the midpoint of the class).

The second set of transitions, which apply only to those whose initial property income is greater than \$250, is virtually identical to the first. The only difference is that the transition



matrices depend on age only, rather than age and income. The idea behind this distinction is that total income is relevant in determining the level of property income a person is likely to achieve in the first instance. That is, the larger the person's total income, the larger his savings are likely to be. And the larger the person's savings, the larger his initial property income. Whether this initial amount of savings is then further built up or drawn down depends mainly on the stage of the life cycle, i.e., age.

#### 6. The Other Money Income Process

Other money income consists of room and board income, alimony, and other small items of income which are difficult to simulate. Because this kind of income is small, and confined to very few people, it is handled in a simple deterministic way. The only persons allowed to have other money income are those who had some in the previous year. The size of the current year amount is increased, however, to reflect any changes in the CPI.

#### 7. Total Income and Output of the Market Income Block

Once all of the components of an individual's income have been determined, his total income for the simulated year is computed by simple summation. This total is then recorded in the individual's state vector.



## 6.2 The Market Income Block Parameters

The following is a brief discussion of all of the parameters that are input to the Market Income Block. Section 6.2.1 discusses the estimation of the employment income parameters. Section 6.2.2 does the same for the property income parameters, and Section 6.2.3 for the retirement income parameters. Appendix E should be consulted, inter alia, for the definition of each set of parameters and the detailed indices embodied in each.

Section 6.2 as a whole is not concerned with estimation procedures as such. For the most part, the Market Income Block parameters do not consist of data estimated by regression techniques, interpolation, and so on. They are rather data (cumulative probabilities for the most part) derived from micro-data files compiled by Statistics Canada, the Unemployment Insurance Commission, and the Department of National Revenue. The discussion that follows is mainly concerned with indicating "why" a given set of data was compiled, rather than some other set. The computer programs, the adjustments to the raw data, and the other details of the compilation process itself are mentioned in only a peripheral way.

### 6.2.1 Employment Income Parameters

The existence of employment income as a major component in a person's state vector presents two problems for POLSIM. The first is the establishment of initial employment income for persons who enter the labour force for the first time during the course of



the simulation. The second problem is that of effecting transitions between income classes for employed persons as they move from year to year. The following discussion details the way in which these problems are handled by the Market Income Block, and the parameters that are necessary to effect the various processes.

1. Income Transitions

(a) General

For purposes of employment income transitions, POLSIM divides the employed population into two mutually exclusive classes. We may designate the people in these classes as "Class A" persons and as "Class B" persons. Class A persons are those for whom the concept of unemployment has no precise meaning, (the self-employed for example), or those who are extremely unlikely to ever experience unemployment. More precisely, Class A persons are males who are either self-employed, or who are employed in managerial, professional, or technical occupations. All other employed persons are Class B persons. These are people who are likely to leave the employed state (for either the non-labour force or the unemployed state) once or many times during their working lives.

The reason for distinguishing between Class A and Class B persons has to do with the obvious fact that employment income depends on the extent to which a person is employed. Employment income for a year is some wage rate multiplied by a period of employment. Annual income can thus change if wages change, if the period of employment changes, or if both of these factors change. The most general approach to income change would be to let both





determinants vary from year to year. But if there exists some group for whom employment changes are not meaningful (they can be assumed to be always fully employed) then it would be both more realistic and simpler in dealing with these people to consider annual wage changes only as the sole determinant of income change.

This is the strategy that is adopted with respect to Class A persons. It is assumed that they never become unemployed and that their income changes are hence determined solely by changes in their annual wage. For Class B persons, on the other hand, it is necessary to examine changes in both weekly wage rates and number of weeks worked.

The Activity Block deals with changes in annual weeks worked. (See Chapter 5.) The major problem faced by the Market Income block is to obtain the income changes that are applicable to the two classes of persons described above. That is, it is necessary to obtain a weekly wage rate transition matrix that is applicable to Class B persons, and an annual employment income transition matrix that is applicable to Class A persons. To obtain these two kinds of matrices we make use of the UIC-DNR data base.

(b) The UIC-DNR Data Base

This data base was originally produced for the Unemployment Insurance Commission to assist in the analysis of proposed new unemployment insurance schemes. It consists of data describing the demographic, financial, and employment characteristics of 2% of the Canadian working population. The data base was compiled



from two main sources: Statistics Canada and the Department of National Revenue. The Statistics Canada files consisted essentially of samples of UIC administrative records, with occupational and industry codes added by Statistics Canada, as well as survey data collected as a joint UIC-Statistics Canada project. The DNR files supplied information from income tax returns.

Persons whose SIN number ends in "14" or "34", a total of approximately 250,000 individuals, comprise the UIC-DNR sample. Creating the data base consisted in a sequential matching of the two basic data files to the sampled individual. Individual files on the insured population, contributions paid and benefits received were obtained from Statistics Canada. These were merged together by matching SIN (if the data existed for the given SIN) to form one file containing all information received from Statistics Canada. At the same time the same sample of individuals was drawn from the DNR files on income tax returns (if a return existed for the given SIN). A final merge was then made combining the SIN master file sample, the Statistics Canada file, and the DNR tax returns file.

The data base thus compiled contained the following data that was relevant for our purposes: demographic characteristics (age, sex, marital status, province), the various components of income a person might have (wages and salaries, business income, etc.), the number of weeks worked, and whether or not the person paid unemployment insurance premiums. These data exist for the 7 years 1965 through 1971 inclusive.



(c) Class B Person's and Weekly Wage Rates

In constructing transition matrices for Class B persons, it is first necessary to identify the subset of the UIC-DNR data base that consists of Class B persons. Roughly speaking, Class B persons are those who are likely to experience unemployment. During the period for which the data exists, 1965-71, the insured population was approximately coterminous with this group: prior to June 1968 the insured population consisted mainly of all wage earners, and all salaried workers earning less than \$5,400; after June 1968 coverage was extended to all wage earners and to all salaried workers earning less than \$7,800. It is thus not unreasonable to derive weekly wage transition matrices from that subset of the data base which contains all persons who had UIC contribution records in this period. The contribution record contains data on the number of weeks worked. The annual wage income for the person is taken from the DNR record. From these it is possible to calculate weekly wage rates. And, by obtaining the person's wage rate in two consecutive years, a transition matrix can be derived.

(d) Class A Persons and Annual Employment Income

If a person has a DNR record on the data base, but has no UIC record in any of the 7 years, then it is reasonable to assume that he is a Class A person. That is, he is a person who is unlikely to ever become unemployed. All self-employed people will fall into this class, as well as salaried workers with high incomes. All wage earners will be excluded. From the DNR record



for these sorts of individuals it is possible to derive employment income (which is the sum of wages, commissions, business net income, professional net income, farming net income, and fishing net income). And since employment income thus defined will exist for several consecutive years, annual employment income transition matrices can be derived.

Annual employment income transitions for all persons are thus handled in a conceptually simple way. For Class A persons, annual transitions determine how their employment income from all sources changes. Class B persons are assumed to have employment income from only one source, wages, and this income is changed by determining transitions in both weeks worked and in weekly wage rates.

(e) Stratification of the Transition Matrices

Stratification is the process wherein a body of data is grouped into a number of disjoint classes. In the case of income transition matrices, for example, it is desirable to have different matrices for different age groups, sex classes, regional classes, and so on. If income transitions are significantly different for different subsets of the population, then stratification will yield far more realistic results.

Unfortunately, the use of stratification variables creates a dilemma. On the one hand, one would like to stratify on all variables that are significant in explaining differences in the data. But if this is done, it usually turns out that the number of cells one ends up with is so large that the resulting





distributions are statistically meaningless. For example, if we are constructing transition matrices with 10 income classes, and if we wish to stratify on age (5 classes, say), marital status (2), province (10), and sex (2), then we will have 20,000 cells into which our data could fall ( $10 \times 10 \times 5 \times 2 \times 10 \times 2$ ). Since the number of records available for computing weekly wage rate transition matrices is approximately 100,000, it is clear that most of these 20,000 cells would contain very few, if any, observations. Obviously, then, we have to restrict the number of stratification variables if we are to have any confidence at all in the statistical veracity of our derived matrices.

The question is, how much restriction is necessary? The table given in Appendix E indicates that to ensure reasonable statistical reliability, we need at least 100 observations in any row of a given matrix. This number is strictly correct if we are dealing with  $2 \times 2$  matrices. For larger matrices, larger numbers of observations would be required. It is not necessary, however, to have this many observations in every row of a given matrix. What we desire is some confidence in the matrix as it will be used. And to achieve this, we require that there be a reasonable number of observations in the rows of the matrices (or even the cells of a given row) that will apply to the great majority of people. The fact that there are very few observations on people moving from very high wage rates to very low wage rates need not worry us very much. What we do want to ensure, however, is that there are a reasonable number of observations around the diagonals of the derived matrices.



The above paragraph abounds in such obscurities as "reasonable", "larger", "some confidence", and so on. The problem is that it is very difficult to come up with some precise statistical measure of how adequate a whole matrix of observations is. The method outlined in Appendix E gives at best a very crude idea of what kinds of numbers to look for. In addition, we do not in fact wish to assess the significance of the matrix as a whole. To reiterate, we do wish to have confidence in the transitions that apply to the large majority of people in the simulation. And this means that we want to look for a "reasonably large" number of observations around the diagonals. Selecting stratification variables is thus as much an art as it is a science, and the rationale set out below reflects this fact.

It was first assumed that the critical stratification variables were age, sex, and region. Transition matrices stratified on these variables were then derived. Five age classes were chosen (14-24, 25-35, 36-45, 45-64, 65+), and combined with the 5 regional classes and 2 sex classes yielded 50 matrices. As expected, some of these matrices were so sparse as to be meaningless. The problem was then to reduce the number of stratifications.

(f) Weekly Wage Rate Transitions (Class B Individuals)

Inspection of the data indicated that all three of the stratification variables were important. That is, the transition matrices were different for different age classes, different sex classes, and different region classes. The differences were what one would expect a priori. Males have higher probabilities of increasing their wages than females. Younger people have higher



probabilities of increases than older people. And people in Ontario, for example, have greater chances of increases than people in the Atlantic provinces.

There is thus no obvious rule by which to eliminate any of the stratifications. All that can be done is to eliminate the least significant ones. There are very large differences between males and females, and so these must be kept. This reduced the choice to either region or age as the variable to be eliminated. Of these two, age is much more critical. People in the 14-24 age group, for example, tend to have many more increases in wages than those in older age classes, where wage rates tend to be more stable. Differences between regions are not nearly so marked. Since it was necessary to eliminate at least one of the stratifications, the choice thus fell to region, as the least critical of the 3 possibilities, and this stratification was in fact eliminated.

Within the age stratification, it was found that the 65+ group was virtually empty. It was therefore decided to aggregate this group with the 45-64 group. The final stratification then consisted of 8 disjoint classes: two sexes and 4 age groups (14-24, 25-35, 36-45, 46+).

(g) Annual Employment Income Transitions (Class A Individuals)

Much the same behaviour was observed when annual wage matrices were compared. The conclusions reached by inspection of annual wage matrices were then applied in the construction



of annual employment income matrices. The reason for this procedure was that the disaggregated wage matrices already existed, whereas the corresponding employment income matrices did not, and it was hence not possible to directly inspect the employment income matrices. The disaggregated income matrices could have been produced, but because of the high cost of doing this it was decided to produce only the final aggregated set. Since wages are the largest component of income, it is unlikely that the conclusions thus reached would have been any different had the complete set of income matrices, stratified on all 3 variables, been produced as well. The only difference between the weekly and annual data was that for the annual employment income transitions (which are to apply to Class A persons) there were very few observations on the 14-24 age group. This group was therefore aggregated with the 24-35 age group.

The final stratification for annual employment income change then consisted of six classes: two sexes and three age groups (14-35, 36-45, and 46+).

(h) Time Series Analysis

The UIC-DNR data base provides 5 observations on weekly wage rate transitions (1965-66 to 1969-70) and 6 observations on the annual wage rate transitions. Given this data one could proceed to examine such questions as the degree to which the matrices vary with time, the extent to which changes can be explained by inflation or other macro-variables, and so on.





Unfortunately, these kinds of analyses are beyond the scope of the present study. What has been done here is to construct transition matrices that represent changes in real income (1969 dollars in the case of weekly matrices; 1970 dollars in the case of annual matrices), and it was assumed that these matrices are time invariant. These matrices were obtained for the final year of data in each case (1969-70 for weekly, 1970-71 for annual). The truth of the time invariance assumption remains an empirical question, one that hopefully can be answered in future work.

(i) Inflation and the Construction of Income Transition Matrices

The procedure adopted was to deflate the higher year's incomes by the change in the consumer price index. Consider, for example, the case of weekly wage rates. The raw data here consisted of a record containing a person's money income in 1969, and his money income in 1970. (As well as the stratification variables, age and sex). The CPI in 1969 was 125.5 (1961 = 100.) and in 1970 it was 129.7. The person's 1970 income was therefore divided by  $129.7/125.5 = 1.033$ , to obtain his 1970 income in 1969 dollars. His 1969 money income and his deflated 1970 money income then determined the cell of that matrix that was to apply to him. All individuals were counted in this way, thus deriving the required matrices. In the same manner for annual transitions, 1971 income was deflated by dividing it by 1.0285. This enabled 1971 incomes to be expressed in 1970 dollars.



The matrices thus derived have a very specific meaning. In the case of the weekly matrices, we can express this meaning as follows:

Let  $P_{ij}(a,s)$  be an element of a given matrix derived as described above;

and let a person's income in any year  $t$  be such that if it is expressed in 1969 dollars it will fall into income class  $i$ ;

Then if the person is in age class  $a$  in year  $t$ , and sex class  $s$ ,  $P_{ij}(a,s)$  is the probability of moving to income class  $j$  in year  $t + 1$ , where income class  $j$  is defined by limits expressed in 1969 dollars.

(j) The Income Transition Program

After a person's state vector has been read, and it has been decided that he will make an income transition, the process proceeds as follows. (The description is for weekly wage rate transitions. With the requisite changes, the process for annual change is identical).

- (i) The person's age and sex class are determined, thus defining the relevant transition matrix.
- (ii) The person's wage is deflated to 1969 dollars.
- (iii) The person's wage class is then determined, and his position in that class, relative to the maximum income in the class, is noted.



- (iv) The income transition (via random sampling of the relevant row of the transition matrix) then defines his wage class in year  $t + 1$ .
- (v) His wage in year  $t + 1$  (in 1969 dollars) is determined by placing him in the same relative position in the new class that obtained for him with respect to the old class in year  $t$ .
- (vi) This wage is then inflated to current dollars in year  $t + 1$ . (The inflation factor in steps (ii) and (vi) is the change in the CPI).

The assumptions embodied in the above procedure are that there exists a time invariant transition matrix explaining real changes in income from year to year, and that money changes in income can be described by a multiplicative change in the determined real incomes. The multiplicative factor is assumed to be defined by changes in the CPI. This same dichotomization of real and money incomes is also assumed with respect to the assignation of initial incomes discussed immediately below.

## 2. Initial Incomes

### (a) The 1971 Data Set

The problem of "Initial Incomes" is very easy to formulate. The POLSIM model will cause, each year, certain individuals to enter the labour force for the first time. These people will



be either students graduating from school, or people such as housewives who leave the non-labour force and obtain employment. These sorts of people will have no income at all (or a very small income obtained from part-time or summer employment) in the year previous to the one in which they became full-time labour force participants. It would therefore be unrealistic to apply an ordinary income transition to these people, since these transitions are meant to apply only to people who are full-time labour force participants in both years being considered. What is necessary is to assign to new entrants an initial income that takes cognizance of their age, sex, education level, and perhaps other factors as well.

The income that is to be assigned would be a weekly wage rate in the case of Class B persons, and a full employment annual income in the case of Class A persons. "Income from employment" in the latter assignment is defined as the sum of wages and salaries, and net income from self-employment.

Data with which one can solve the initial income problem is far from ideal. What one would like to have is a joint distribution of the incomes of first-time labour force participants cross-classified by all of the relevant individual characteristics. Unfortunately, such data simply doesn't exist. One is thus forced to examine all of the data that does exist, and piece together as large a joint distribution as possible, on the basis of reasonable assumptions.

Richard Arnott has carried out this exercise, as part of an Education Finance Study undertaken by the Institute for Policy





Analysis at the University of Toronto. He calculated a distribution of 1971 mean full employment incomes cross-classified by age (57 categories), sex (2), education (10), marital status (3), and province (10). More specifically, his data is cross-classified as follows:

- (i) 57 years of age (ages 14 through 70)
- (ii) sex
- (iii) 10 education classes:

- no schooling
- some elementary
- elementary completed
- some high school
- high school completed
- some university
- community college graduate
- Bachelor's degree
- Master's degree
- PHD

- (iv) Marital Status

- single
- married
- other

- (v) The ten provinces



This is a very extensive and useful set of data, particularly since it breaks education down into such fine categories. The only difficulty with it is that it applies only to the year 1971. Since the POLSIM model will operate for years subsequent to 1971, it is necessary to adjust the data to reflect inflation and economic growth. The whole procedure for assigning an initial income to a given person will then consist of the following steps.

- (i) An individual's 1971 mean full employment income will be calculated on the basis of the Arnott data;
- (ii) This income will then be adjusted to account for inflation and economic growth;
- (iii) If a person is designated as Class B, then this income will be divided by 52 to give a weekly wage rate.

It remains now to explain step (ii) above.

(b) Inflation and Growth

Conceptually, this problem is quite straightforward. We begin with the person's 1971 income in 1971 dollars. What we want is his 1975 income (say) in 1975 dollars. We proceed as follows:



- (i) Let  $g(a,e,s)$  be an exponential growth factor defining the yearly growth in real incomes for someone in a particular age, education, and sex class.

Then if  $Y_R^{71}$  is his real income in 1971, his real income (1971 dollars) in year  $1971 + t$  will be

$$Y_R^{71+t} = Y_R^{71} e^{gt}.$$

- (ii) We now have the person's real income in the required year, but expressed in 1971 dollars. If  $CPI(t)$  is the consumer price index in year  $1971 + t$ , the person's money income in year  $1971 + t$  will be

$$Y_m^{71+t} = Y_R^{71+t} \times \frac{CPI(t)}{CPI(0)}$$

(c) Estimation of Growth Factors

The only problem we have yet to deal with is the calculation of the growth factors. And again, this is a relatively simple problem. From the 1967 and 1971 surveys of Consumer Finance, we can obtain mean full employment net employment incomes cross-classified as follows:

- (i) age

14-17

18-21

22-25

26-29



30-33

34-37

38-41

42-45

46

(ii) sex

(iii) education

less than grade 9

less than grade 12

grade 12 or 13

some Univ. or CAAT

CAAT or Univ. Grad.

Post Graduate

We thus have  $Y_{67}^M(a,s,e)$  and  $Y_{71}^M(a,s,e)$  where  $Y_t^M(a,s,e)$  is the mean money income of all persons in a given age-sex-education class in year  $t$ .

Now define real incomes as follows:

$$Y_{71}^R = Y_{71}^M$$

$$Y_{67}^R = Y_{67}^M \times \frac{CPI(71)}{CPI(67)}$$

The growth factor for the given class is then defined by

$$Y_{71}^R = Y_{67}^R e^{4g}$$

$$\text{or } g = \frac{\ln Y_{71}^R - \ln Y_{67}^R}{4}$$





### 6.2.2 Property Income Parameters

The property income transition matrices were also derived from the UIC-DNR data base. The DNR record on this file contained data both on an individuals dividend income and on his income in the form of interest and returns from other investments. This data exists for the three years 1969, 1970 and 1971. It was thus possible to obtain two observations on single year dividend transition matrices and single year interest transition matrices.

#### 1. Aggregation of the Time Series Data

Two observations are not enough data to make reasonable inferences concerning such underlying determinants of property income transitions as the state of the business cycle, the rate of inflation, and so on. It was therefore decided to aggregate the two observations in order to reduce small sample error and randomness in the data. The aggregation process consisted simply of adding the number of counts in any given cell of the matrix for 1969-70 to the number of counts in the same cell for the 1970-71 matrix. We can make the assumptions inherent in this process explicit: Let  $(p_1, p_2, \dots, p_n)$  be any row of the 1969-70 matrix and let  $N$  be the total number of counts in that row. Then  $(Np_1, Np_2, \dots, Np_n)$  is the vector of counts for that particular row. Similarly, let  $(q_1, q_2, \dots, q_n)$  be the same row of the 1970-71 matrix, and let  $M$  be the total number of counts in that row. Then  $(Mq_1, Mq_2, \dots, Mq_n)$  is the vector of counts for that row. Let  $(S_1, S_2, \dots, S_n)$  be the derived row of the aggregate transition matrix.



$$\text{Then } S_i = \frac{Np_i + Mq_i}{N + M} = \frac{N}{N + M} p_i + \frac{M}{N + M} q_i$$

That is,  $S_i$  is just the weighted average of the original probabilities, where the weight depends on the relative number of counts in the original rows. If  $N$  is very small in comparison with  $M$ , for example, then very little weight will be attached to the 1969-70 probability. And this is what we wish, since if  $N$  is small it is likely that there will be large errors in the  $p_i$ 's as compared with the errors in the  $q_i$ 's.

## 2. Stratification

The matrices were originally made dependent on both age and income. There were 4 age classes (14-35, 35-50, 50-65, 65+) and 3 income classes (0-7k, 7k-15k, 15k+).

### (a) Age Effect

It was felt that money property income transitions would almost certainly depend on age, although just what these effects would be was not entirely obvious. Differences between the first two age classes were felt to be ambiguous. On the one hand, people in their early earning years might hold their savings in the form of financial assets such as stocks and bonds, and then liquidate these during the "middle" earning period to acquire real assets such as houses, etc. This would imply decreased probability of raising one's money property income in the second age period. On the other hand, since income is correlated with age, it was felt that many people would not even begin to invest in financial assets



until they had reached the second age bracket. Having purchased most of the "necessities" of life during their early earning years, and having experienced an increasing income during these years, they would not be in a position to invest in stocks, bonds, etc. This would imply increased probabilities of raising one's money property income in the 35-49 age period, especially for those initially in the zero or very low property income classes.

The data reflected this ambiguity. If we examine Table 6.1, which illustrates the way that the interest income probabilities behave, we can see that for the low initial interest classes (less than \$750), the probability of moving into a higher interest income class generally increases as people move into the second age bracket. For people in higher initial interest classes, on the other hand, the probabilities decrease as people reach the higher age class. Much the same behavior is exemplified in the dividend transition matrices as well. As explained above, these results are as one might expect. Whether the probabilities increase or decrease as one moves into the second age bracket is likely to depend on how much property income the person had to begin with.

In comparing the third age class with the second, it was thought that the probabilities should be higher in the former for all income classes. That is, people in the 50-64 age class should be able to increase their property incomes more frequently than people in the 35-49 age class. The data indicated, quite generally, that this was indeed the case. (Table 6.1 illustrates this behaviour for interest incomes. The dividend matrices are again similar.)



Table 6.1

Probability of interest income increasing or remaining constant for persons in different age classes and different initial interest income classes

Initial Interest Class	Age Class			
	14-34	35-49	50-64	65+
0	1.000	1.000	1.000	1.000
1-250	.6942	.7914	.8369	.8600
251-500	.6189	.6544	.7069	.7386
501-750	.5783	.6292	.6778	.6884
751-1K	.6041	.5876	.6601	.6570
1K-2K	.7415	.7061	.7715	.7814
2K-3K	.6714	.6746	.7368	.7562
3K-4K	.7368	.6623	.7297	.7363
4K-5K	.6875	.5729	.6761	.7490
5K-6K	.5000	.7432	.6145	.7455
6K-7K	.6000	.5143	.6800	.7200
7K-8K	.5333	.4828	.6000	.7872
8K-	.7037	.7233	.7436	.9091

Table 6.2

Probability of interest income increasing or remaining constant for persons in different total income classes and different initial interest income classes

Initial Interest Class	Total Income Class		
	0-7000	7K-15K	15K+
0	1.000	1.000	1.000
1-250	.7545	.7869	.8535
257-500	.6811	.6833	.6732
501-750	.6579	.6452	.6604
751-1K	.6327	.6595	.5792
1K-2K	.7661	.7486	.7277
2K-3K	.7189	.7505	.7295
3K-4K	.7343	.7052	.7040
4K-5K	.7069	.6895	.6400
5K-6K	.6725	.7437	.5900
6K-7K	.6000	.7329	.5781
7K-8K	.3333	.7014	.6034
8K-	.5000	.8138	.8055





Comparisons between the 50-64 group and the 65 and over group were expected to involve ambiguities. On the one hand, retired people might run down their liquid assets to make up for their lost employment income. Or, they might convert real assets such as houses into assets yielding a monetary return. In the former case, property income would go down; in the latter case, it would go up. The data indicates that the latter tendency is strongest. The probability of remaining in the same position or of improving one's property income increases as people move into the 65 and over age group. (See Table 6.1)

(b) Total Income Effect

Anticipated effects in the case of income stratifications are less clear. It was felt that income would be significant in explaining the level of a person's property income. But it was not at all obvious that income was relevant in explaining changes in property income. The data tended to reflect these presumptions. The first two rows of the transition matrices varied quite strongly with income. As income increased, the probability of moving from the 0\$ class or the 1-250\$ class to a higher property income class increased quite significantly. But for people in higher initial property income classes, the probability of doing better did not change much at all as income increased. (See the probabilities for interest income in Table 6.2. The Dividend matrices reflect the same behaviour.)

This was a fortuitous but fortunate result. When the transition matrices were stratified on both age and income, there were very few observations in the last 11 rows. This meant that



a large sampling error would be introduced into the probabilities in these rows if both stratifications were kept. Because the observations in these rows did not depend significantly on income (as was inferred by examining matrices aggregated over age) it was possible to aggregate these rows over income. The first two rows, however, in which income was important, did contain enough observations to make the complete age-income stratification possible. The net result, then, for both interest and dividends, was two sets of matrices. The first, containing 2 rows and 13 columns, were stratified on age and income. The second, containing 13 columns and the remaining 11 rows, were stratified on age alone.

#### 6.2.3 Retirement Income Parameters

The retirement income process in POLSIM consists simply of the assignment of an initial pension to people when they first retire, provided they are eligible for a private pension. A model of how this initial retirement income will change over time has not been considered in this version of POLSIM. It would involve consideration of such factors as the type of pension plan a person had, whether or not he is a holder of an annuity, the kind of annuity, changes in his plan that are consequent on the death of his spouse, and so on. Such a model would hopefully be incorporated into a later version of the Market Income Block. For the present, retirement incomes, once established, are assumed to be constant through time.

The initial pension distribution was taken from the Statistics Canada Publication, "Survey of Pension Plan Coverage 1965", Catalogue number 74-506, Table 35. The table gives the



distribution of annual pensions as a percentage of final annual earnings for pension plan members who retired during the year ending December 31, 1965. The distribution is by sex and annual earnings group.

### 6.3 Validation of the Market Income Block

#### 6.3.1 Introduction

The parameters of the Market Income Block were validated by constructing expected income distributions on the basis of the transition matrices that are input to the block. The approach was to first compile sets of income distributions from the 1967 Survey of Consumer Finance. Each distribution corresponded to a particular transition matrix. (For example, annual wages of Class A persons, ages 14-35). These distributions were then multiplied by the fourth power of the relevant transition matrix, to produce the expected 1971 distributions. The actual 1971 distributions were then constructed from the 1971 Survey of Consumer Finance, and these were compared with the expected 1971 distributions. The results of these comparisons are presented in the graphs on the following pages.

The above procedure is approximately equivalent to doing a four year simulation. It is not, however, identical, since an actual simulation would account for new entrants into the labour force, demographic changes in the population, retirements, and so on. The expected value approach, on the other hand, has the advantages of cheapness and the absence of Monte Carlo error.



For this reason the method of expected values, as outlined above, was used as preliminary validation. When the results of the 1973 Survey of Consumer Finance become available, it will be possible to do a simulation validation from the 1971 initial year tape. The two validations, together, will permit the estimation of the extent of Monte Carlo error. Finally, a full simulation using the entire model (see Ch. 7) will provide the ultimate test (since it accounts for new entrants, etc., as well as illustrating the extent of the Monte Carlo errors). The expected value approach gives a good test of the parameters per se, and indicates how well they can be expected to perform over a four year period.

#### 6.3.2 Annual Wage Transitions

The first three graphs in figure 6.2 compare the actual 1971 annual income distributions with the corresponding expected distributions. Overall, the correspondence is quite good. The graph for the 14-35 age group shows a shift in the expected distributions towards higher incomes. To a lesser extent this is also true for the 36-45 age group. The reason for these shifts is that the transition matrices for Class A persons were estimated from 1970-71 data. Real growth in wages and salaries for that year was 6.9%, whereas it averaged only 5.8% over the 1967-71 period. Consequently one would expect the transition matrices to yield a higher overall distribution for the 14-35 age group, which is the age class with the highest growth in income. To a lesser extent, one would expect the same phenomenon for the 36-45 age group as well. Since this is exactly what happens, and since the two distributions are otherwise very similar, we have reasonable confidence in this set of matrices.





ANNUAL WAGES, 1971 CLASS A, AGES 14-35

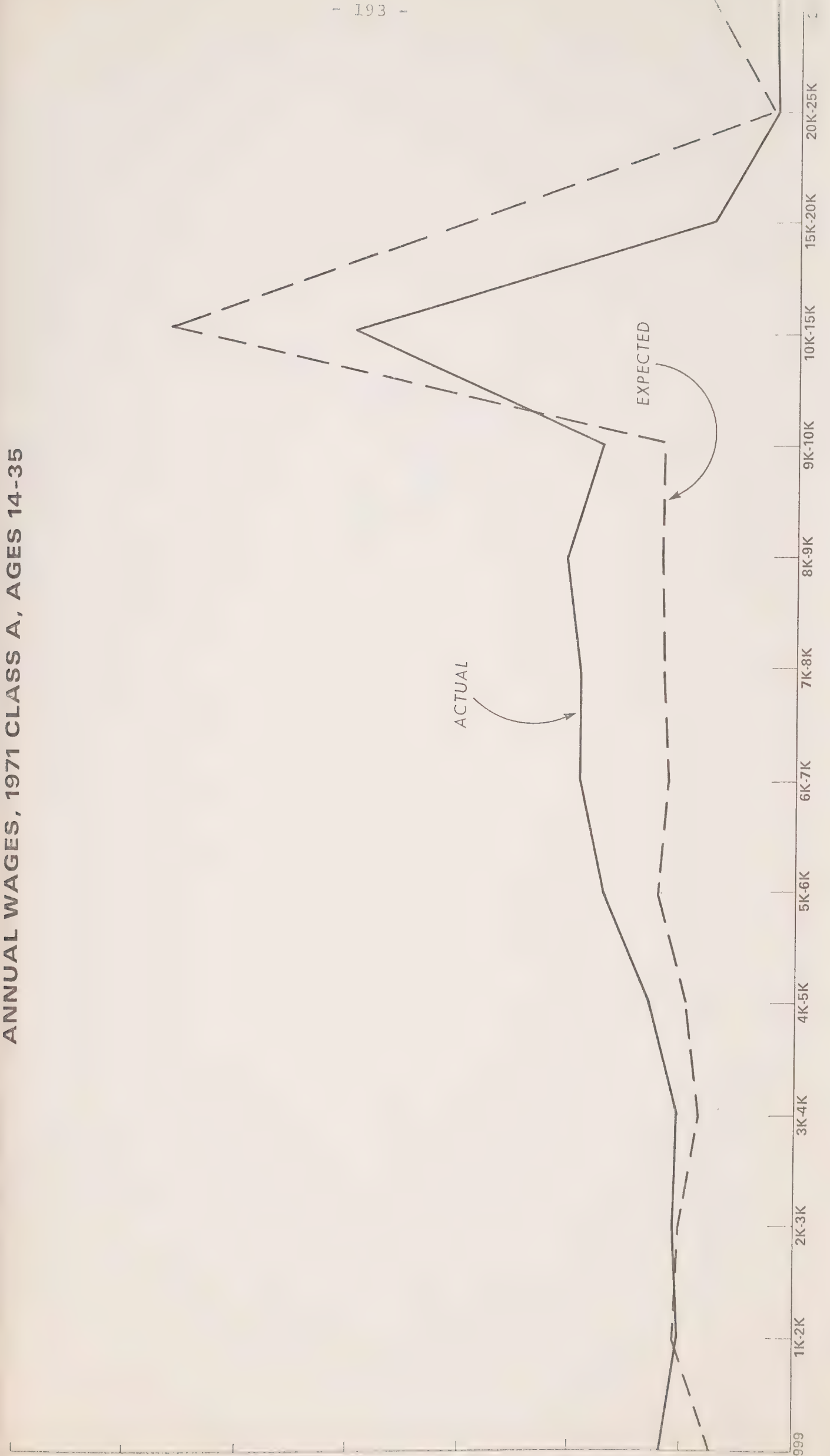
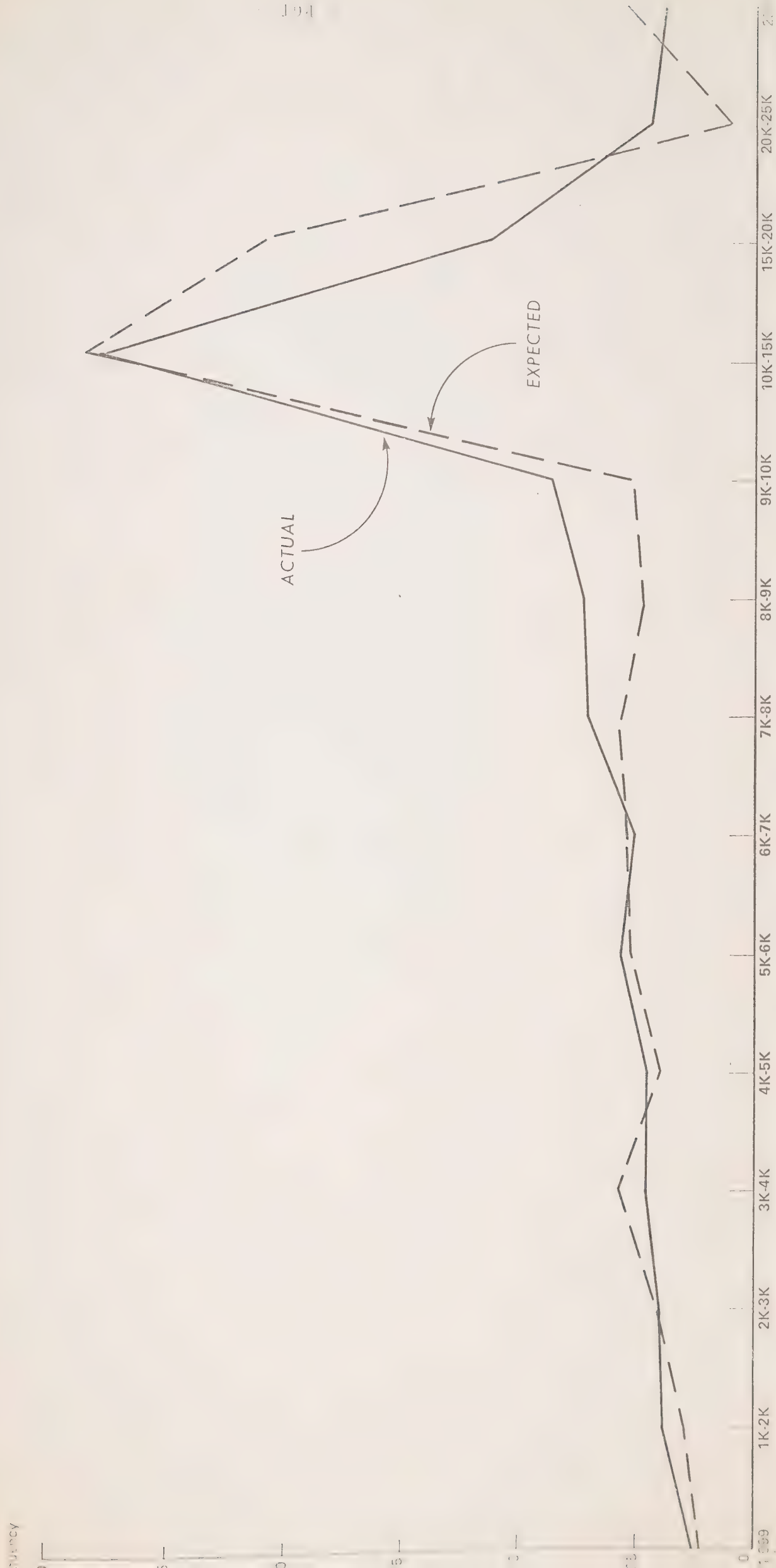




Fig 6.2.2.b

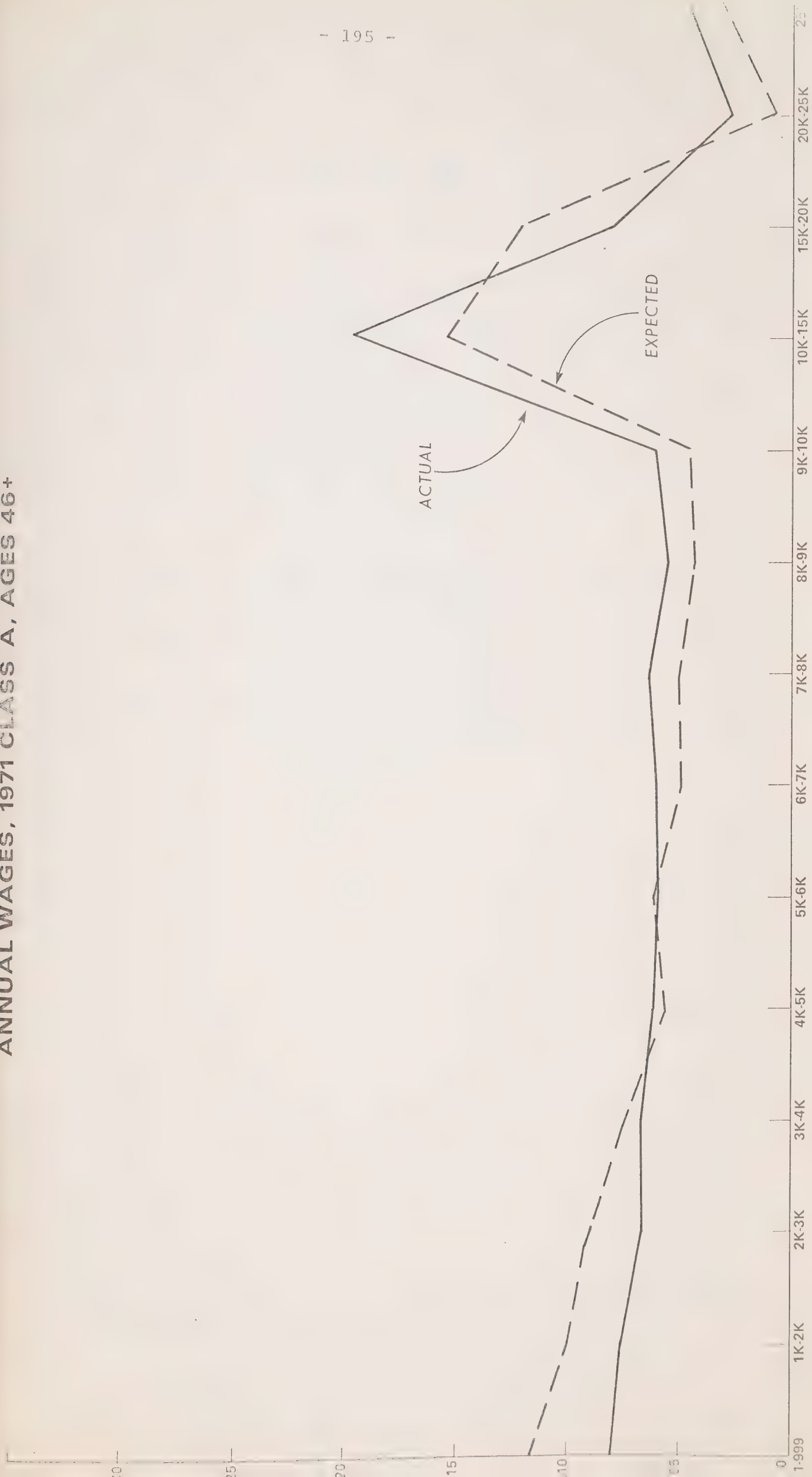
ANNUAL WAGES, 1971 CLASS A, AGES 36-45





# ANNUAL WAGES, 1971 CLASS A, AGES 46+

Frequency





### 6.3.3 Weekly Wage Rate Transitions

The graphs in figures 6.2d - 6.2k illustrate the weekly wage rate comparisons. Again the correspondence between expected and actual is quite good, although two dissimilarities should be noted. The first is a result of the fact that the real growth rate implicit in the transition matrices is lower than the growth rate that actually obtained over the 1967-71 period\*, and the second is caused by the absence of new entrants in the calculation of the expected values.

The weekly transition matrices were calculated from 1969-70 data. In that year the real growth in wages and salaries was 4.8%. In the 1967-71 period the average growth rate was 5.8%. Consequently one would expect the simulated 1971 distributions to be shifted slightly to the left, and in general it can be observed that this is indeed the case.

The only exception to this leftward shift arises in the case of persons in the 14-35 age group. The shift for this group is counterbalanced by the distortion arising from the absence of new entrants. Since the calculation of the expected 1971 distribution is based solely on transitions made by the 1967 working population, and hence takes no account whatever of new entrants to the labour force over the 67-71 period, the 1971 expected distribution is representative of a mature working population: one that has been working for at least 4 years. This population will obviously have a higher average income than the corresponding actual population which includes new entrants. The effect of this

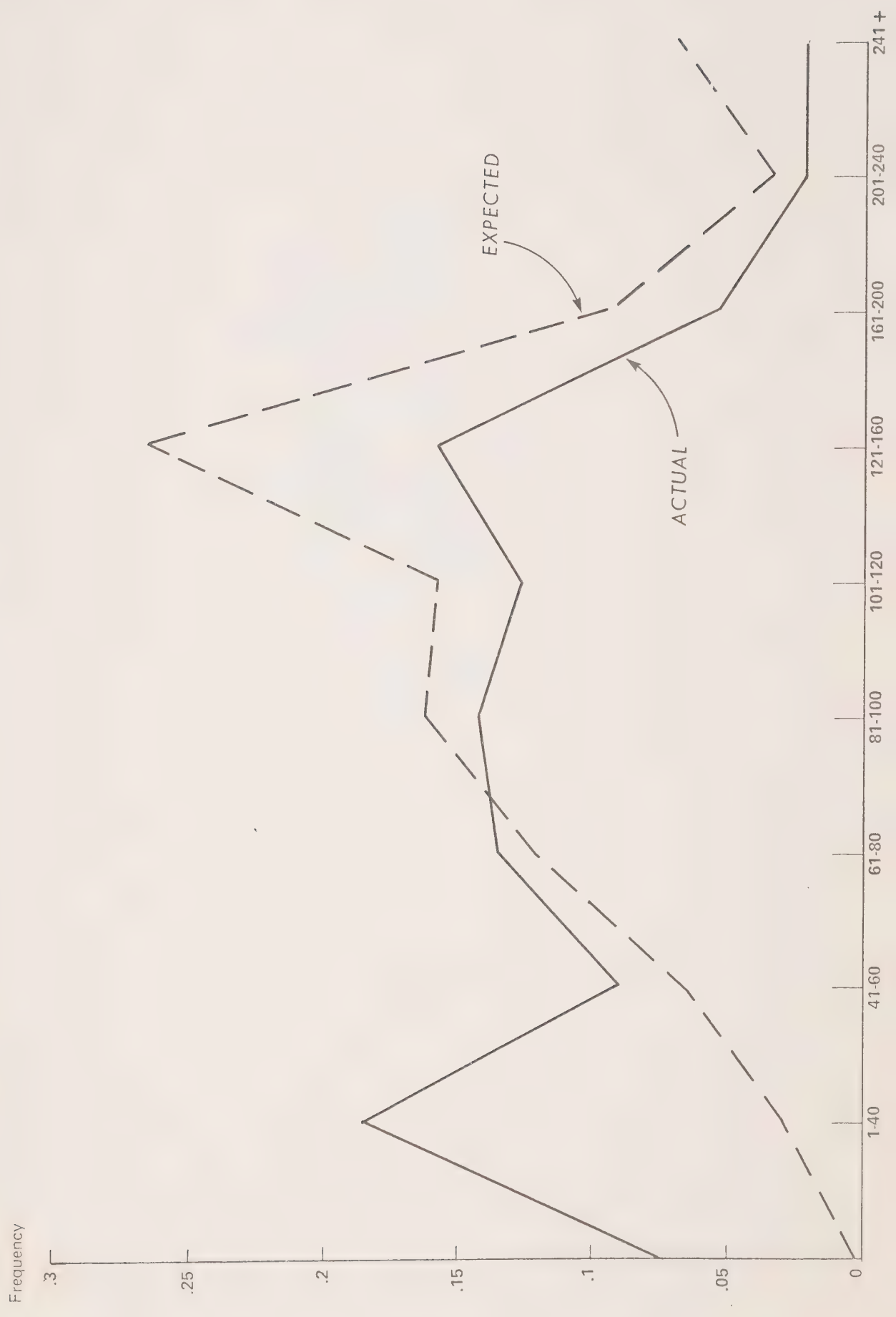
---

\*This error can be corrected, in an actual simulation, by varying the exogenous rates of inflation.



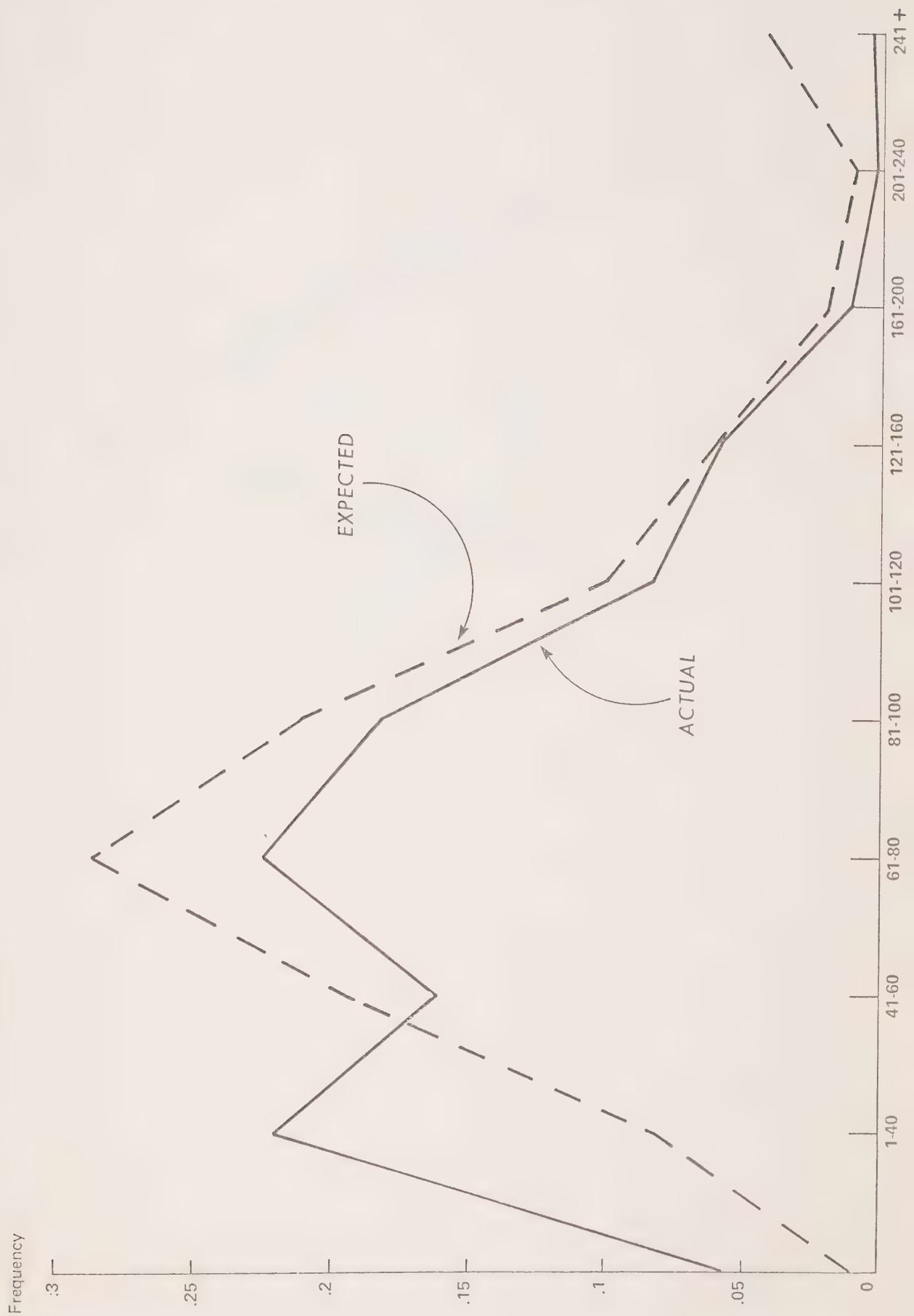


WEEKLY WAGE RATES, 1971  
CLASS B MALES, AGES 14-35



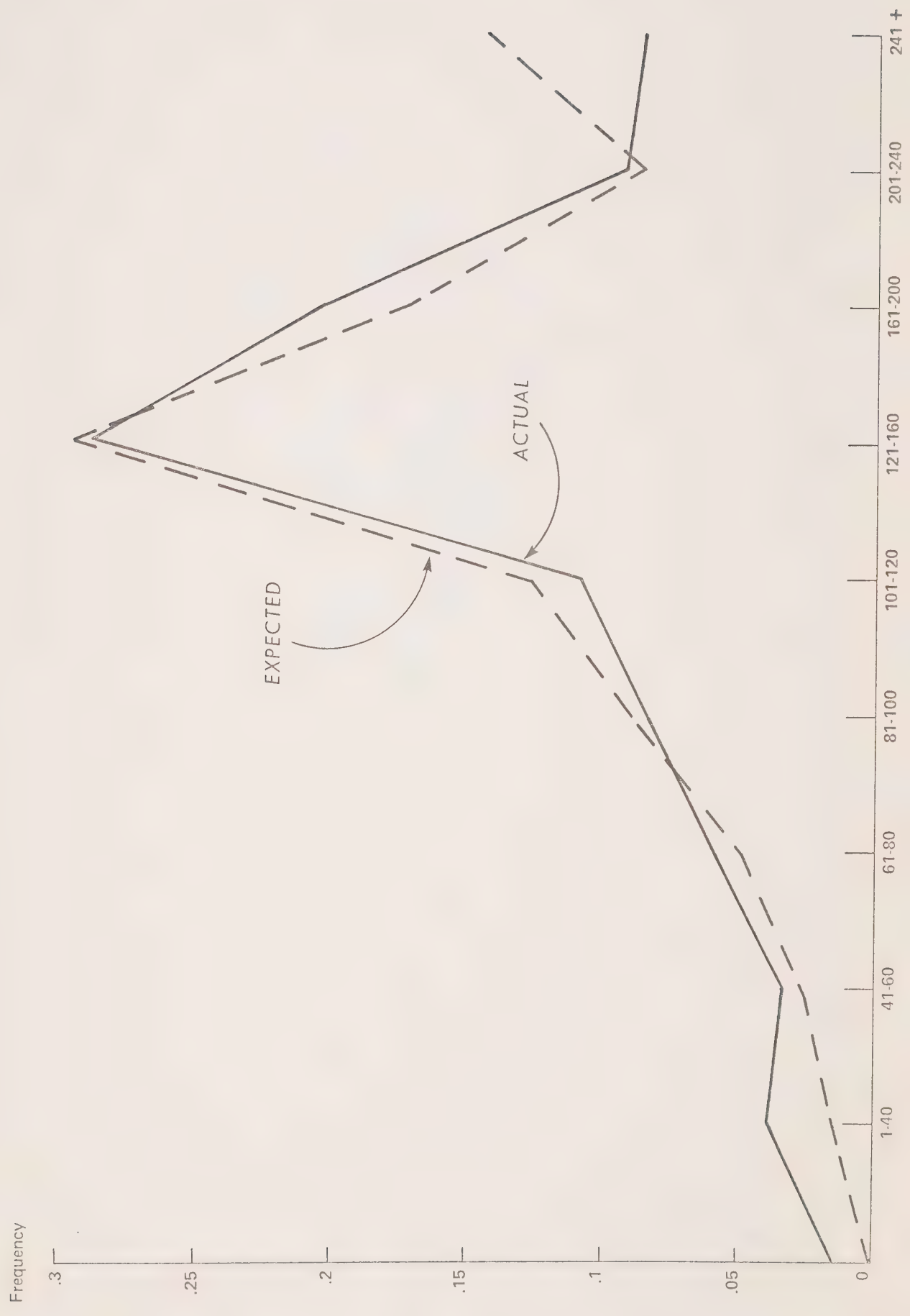


WEEKLY WAGE RATES, 1971  
CLASS B FEMALES, AGES 14-35

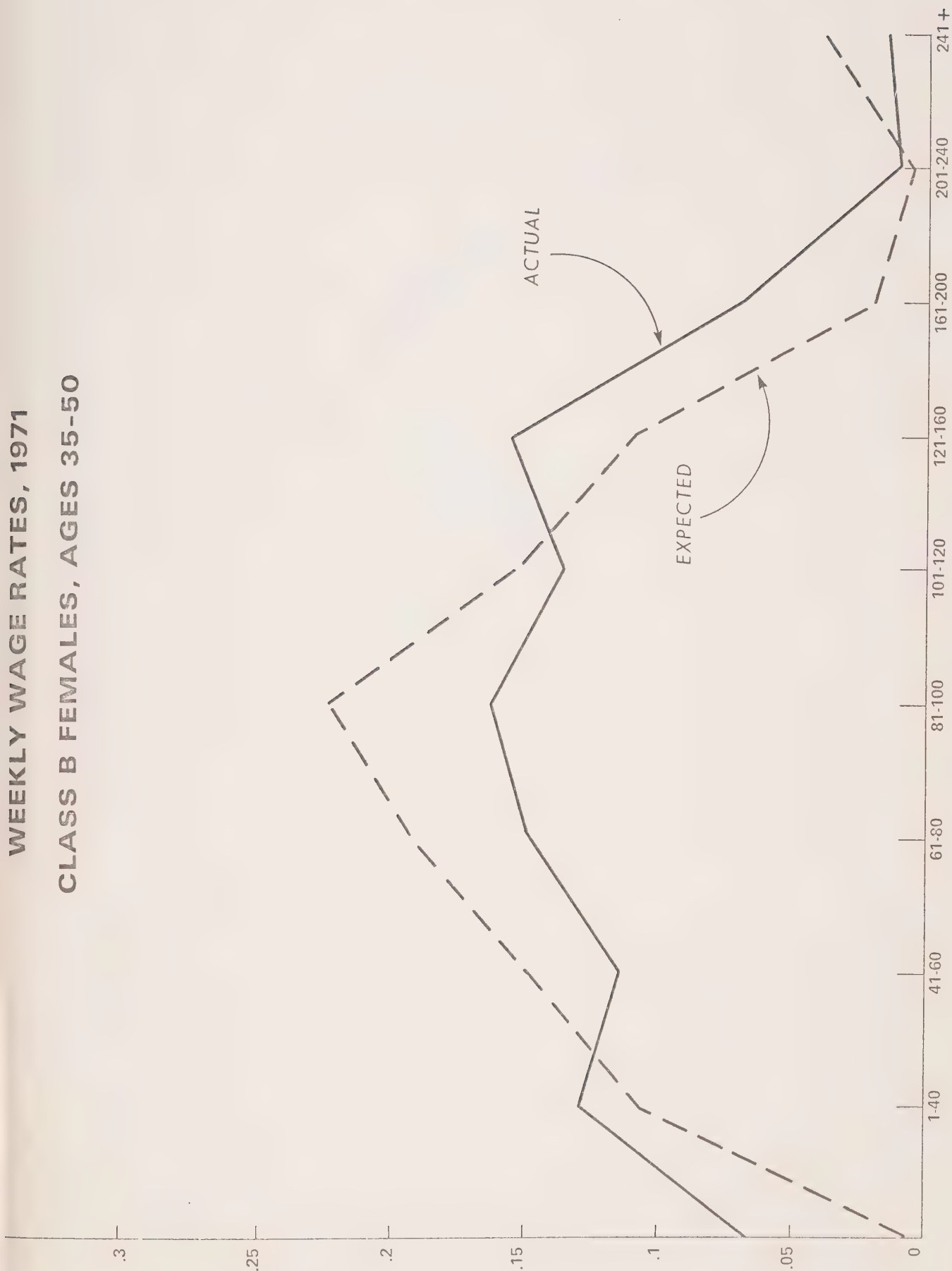




WEEKLY WAGE RATES, 1971  
CLASS B MALES, AGES 35-50



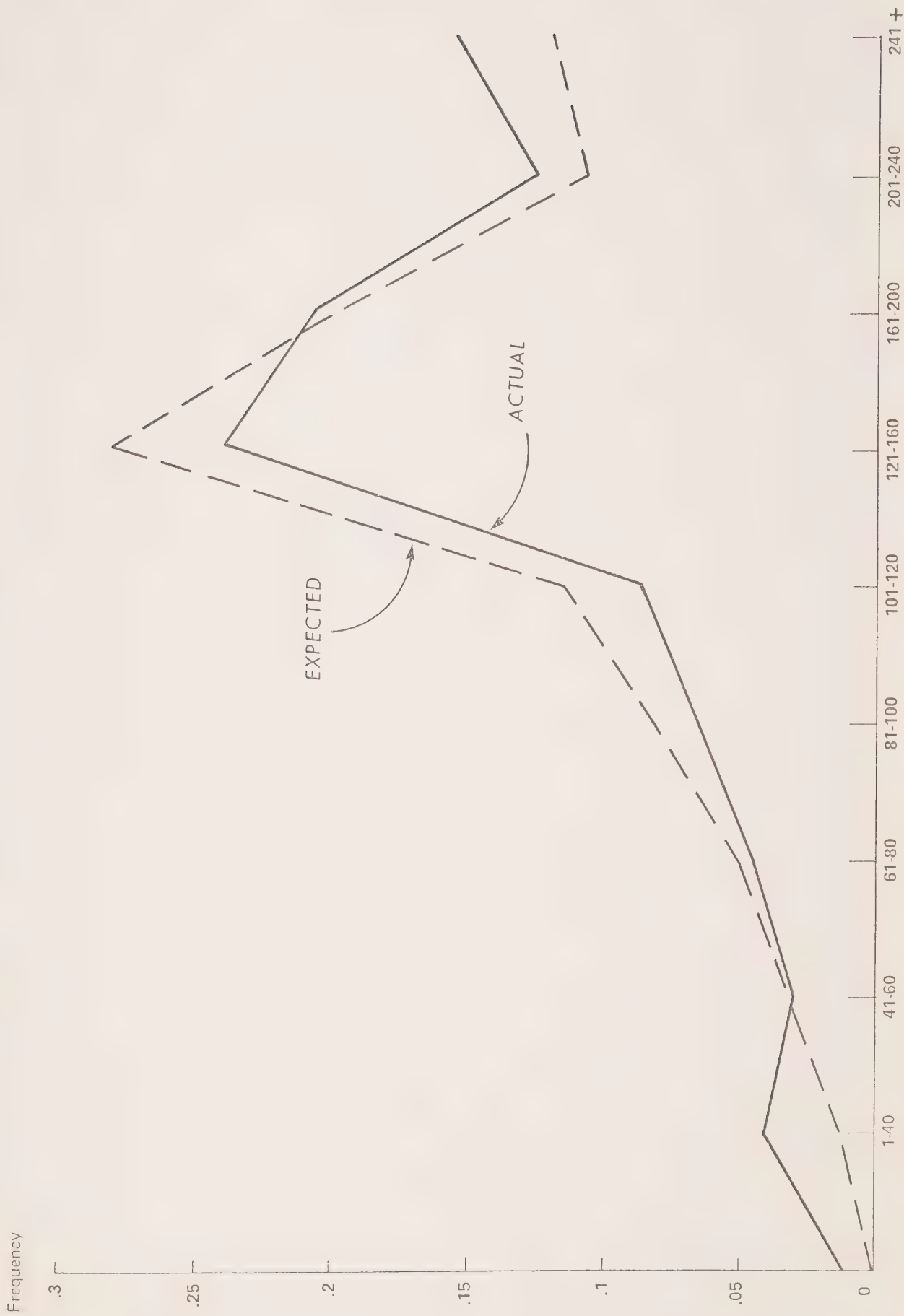




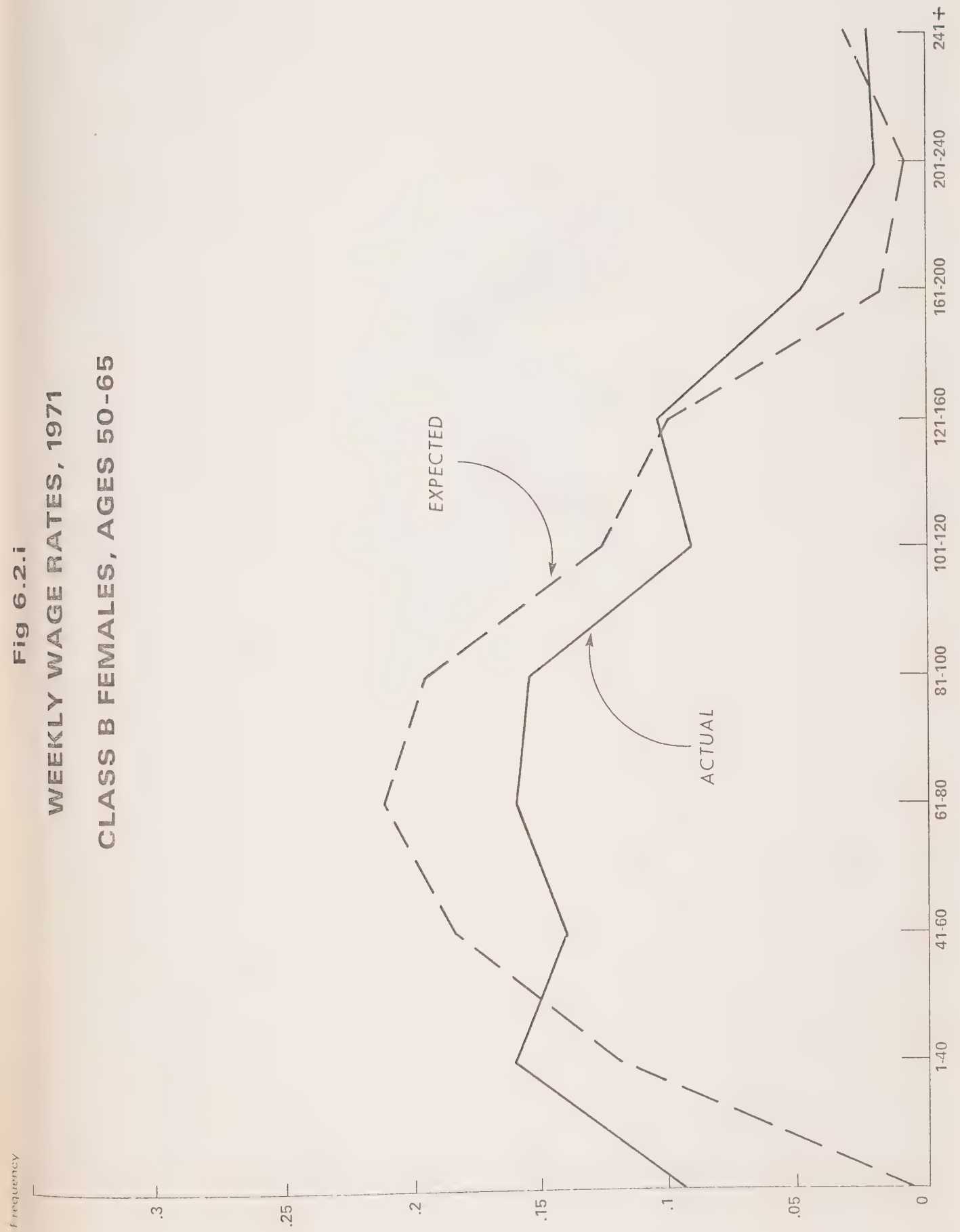




WEEKLY WAGE RATES, 1971  
CLASS B MALES, AGES 50-65









**Fig 6.2.j**  
**WEEKLY WAGE RATES, 1971**  
**CLASS B MALES AGES 65+**

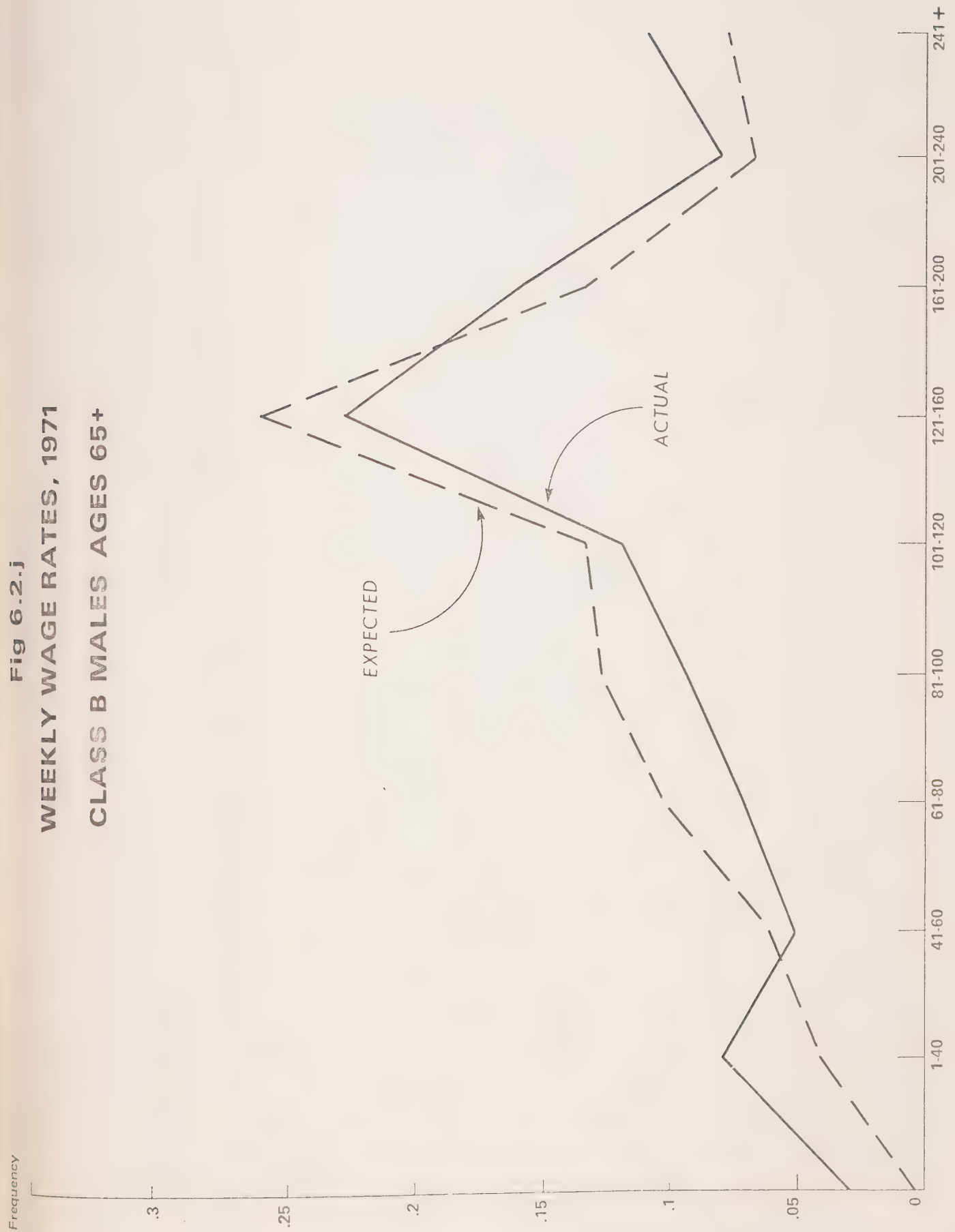
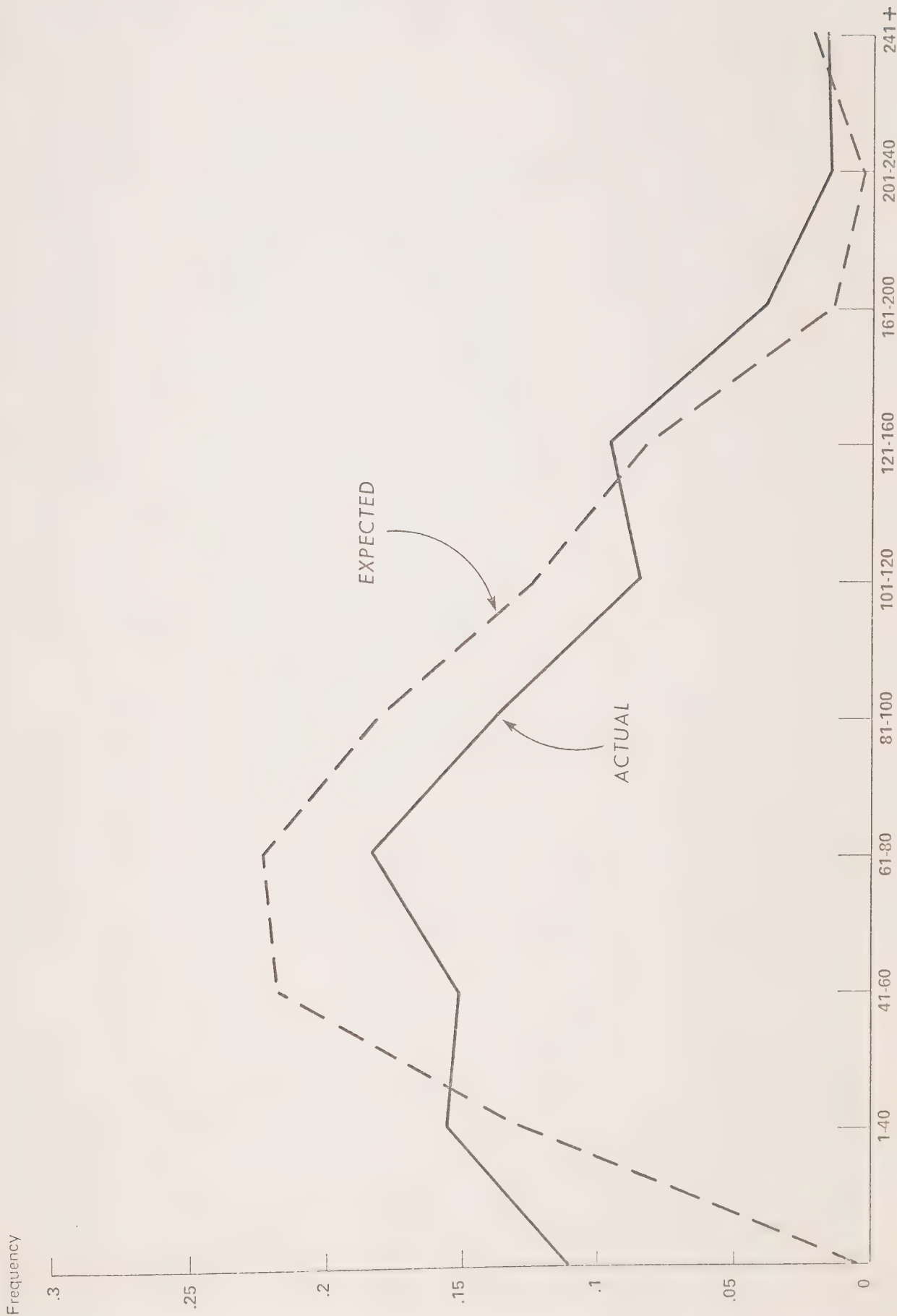




Fig 6.2.k  
WEEKLY WAGE RATES, 1971  
CLASS B FEMALES, AGES 65+







will be to make the expected frequencies too low for low income classes (where the majority of new entrants fall), and consequently too high for high income classes.

Taking these two sources of error into account, the comparison is quite good. The simulation model will eliminate the problem of new entrants, thus producing better results than are indicated by the comparison of the expected and actual distributions. The problem of an implied rate of real growth within the transition matrices remains. But as mentioned above, this can be corrected for by adjusting the exogenous rates of inflation. A future version of the model could possibly generalize the transition matrices to account for time variance.

#### 6.3.4 Property Income Transitions

The last four graphs in figure 6.2 refer to property incomes. Since the 1967 Survey of Consumer Finance did not collect data on dividends, only the "interest and other investment income" component of the property income transition matrices was tested. It is fair to assume that the dividend transitions would reflect much the same behaviour as the interest transitions.

The four graphs indicate that although the expected distributions follow the actual distributions quite closely, there is a general tendency for the transition matrices to overestimate interest income. This is especially evident if we examine the zero income class. For all four age groups, the expected frequencies in the zero income class are less than the actual frequencies.



Fig 6.2.I  
INTEREST INCOME AGE 14-35, 1971

Frequency





INTEREST INCOME, 1971, AGE 35-50





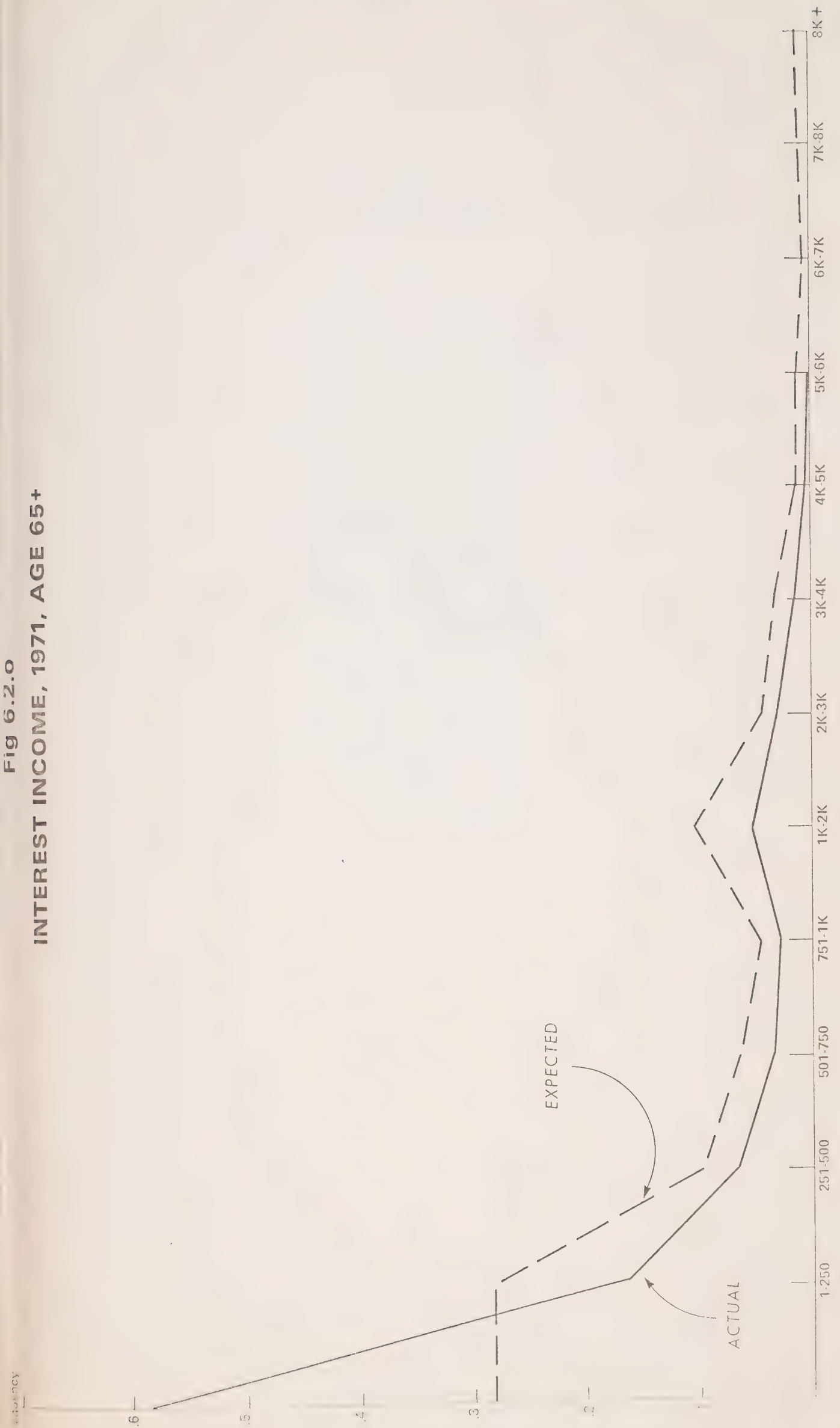
INTEREST INCOME, 1971, AGE 50-65







Fig 6.2.0  
INTEREST INCOME, 1971, AGE 65+





And as a consequence, the expected frequencies for the positive income classes are all slightly higher than the respective actual frequencies.

The reason for this is that the data from which the transition matrices were estimated includes only persons who file tax returns or who make unemployment insurance contributions in the year for which the data applies (1969, 70, or 71). Those persons who would not fall into either of these two categories would be people who would be expected to have zero property incomes. Consequently the transition matrices will be biased in favour of inducing higher property incomes. This is a limitation that cannot be corrected by existing data. The Department of National Revenue is currently initiating a study that would bring non-filers into their data base. Once this has been accomplished, a re-estimation of the transition matrices should eliminate the above bias. In any case, as the graphs indicate, the error is not particularly serious.



## 7. Full Model Simulation: 1967 to 1971

### 7.1 Introduction

The complete POLSIM model was tested by simulating the four year period from 1967 to 1971. The initial year population for the simulation consisted of 397,960 individual records derived from the 1967 Survey of Consumer Finance. The input parameters to the simulation (unemployment rates, total number of immigrants, and rates of wage inflation) were the actual values that obtained over the four year period.

The computer program was set up so that all of the separate blocks would follow one another automatically. The simulation then proceeded one year at a time. For any given year, the Immigration block was run first. The output file of new immigrants was then merged with the previous year final output file (or the initial year population in the case of the 1967-68 simulation). This new file was then input to the Demographic Block which in turn passed its output file to the Activity Block for processing of the Activity variables. The Market Income Block then updated the income components of the individual state vectors thus completing a one year simulation. This whole procedure was carried out four times, the final output being a synthetic population representative of the 1971 population of Canada.

Once the simulated 1971 population had been produced, it was possible to compare it to another estimate of the 1971 population as measured by the 1971 Survey of Consumer Finance. This comparison was then a measure of the adequacy of the POLSIM model for simulations of four years or less.



The details of the simulation from 1967 to 1971 are presented in section 7.2 below. The comparisons between the simulated 1971 population and the other estimate of the 1971 population are then discussed in section 7.3.

## 7.2 Simulation Results: 1967-1971

The results of the simulation of new immigrants over the four year period are summarized in Table 7.1. It can be seen that the model slightly underestimates the total number of immigrants, even though the actual number is an exogenous input. The reason for this is that children are created stochastically in the Immigration Block, and that slight adjustments have to be made to the number of married women so as to equate them with the number of married men. These adjustments have been discussed extensively in Chapter 3.

The results of the Demographic Block processes are summarized in Tables 7.2 to 7.5. The aggregate totals produced by the Demographic Block are compared with actual figures derived from Vital Statistics. In general, the model performs very well. The tables indicate that the model tends to underestimate the number of live births, deaths and divorces while overestimating the number of marriages.

The analysis of the errors inherent in the Demographic Block processes is presented in Chapter 4. This analysis is, however, related to particular groups with common probability of success, and is not directly applicable to the case of





Table 7.1

Comparison of simulated and actual data\*

Total number of immigrants: 1968-71

	<u>Simulated</u>	<u>Actual</u>
1968	180,250	183,974
1969	157,850	161,531
1970	144,200	147,713
1971	118,450	121,900

\* Source: Immigration Statistics, 1967-1971.



aggregates. Any error in an aggregate figure, such as the total number of births or deaths, can be attributed to the following causes: (1) errors in the initial population at risk; (2) errors in the probability parameters; (3) simulation errors; and (4) the additive or cancelling effect of the above three types of errors.

In Chapter 4 we have analyzed the first three types of errors. However, the fourth type is much more complex, since it is related to the whole spectrum of the population groups at risk. For this reason we will not give a complete analysis of the errors of each process. Rather, we will point out informally the reasons for any deviation of our simulated aggregates from the actual ones.

The results of the birth process are given in table 7.2, and it can be seen that the simulation begins with a fairly large error in 1968 which then declines until 1971 when the results are almost perfect. Part of the error in the early years (approximately 15,000 live births) can be attributed to initial population errors. In particular, the initial population is largely underestimated for women in the 20-24 age group, and since this is the prime child-bearing age, a large underestimate in births will naturally ensue. The underestimate in births declines as the simulation progresses for two reasons. First the model uses stationary fertility probabilities estimated for the latest year. The birth rate declined over the period 1967-1971, and since this was not reflected in the probabilities, one would expect any underestimate to decline as time goes forward. Second, the underestimate will also decline as time progresses



because the effect of the initial population error will decrease. In the initial population, females in the 15-19 age group are only slightly underestimated (as compared with the large underestimate in the 20-24 age group). Therefore in each succeeding year, the 20-24 female age group will more closely approximate the true population in that age group, and hence the underestimate in births which results from an underestimate in the 20-24 population will progressively be eliminated. In examining the regional results, table 7.2 indicates that there is no evidence that fertility probabilities should be regionalized.

The divorce process results are given in table 7.4, and are much better than expected. It is known that Canada passed through a transient period during the time of the simulation, insofar as the incidence of divorces is concerned, due to a change in the divorce law in 1969. This made the simulation of divorce quite difficult, and hence the results are on the whole quite pleasing. It is obvious, however, that the regionalization of the divorce probabilities should be seriously considered. In 1970, for example, the model significantly overestimates the number of divorces in Quebec while underestimating the number in Ontario. It is clear that the number of divorces in these two processes cannot be considered to be outcomes of the same stochastic process.

Table 7.3 indicates that marriages are slightly overestimated. This can be explained by the fact that any individual of marital status "other" is, in the model, eligible for marriage. We recall that the state-variable marital-status can attain three state-codes, i.e. single,



Table 7.2

Comparison of simulated and actual data \*

Live Births by Region: 1968-71

		Atlantic	Quebec	Ontario	Prairies	B.C.	Canada
1968	Simulated	26,750	81,600	104,000	44,300	28,350	285,000
	Actual	40,306	96,622	126,257	65,770	33,687	364,310
1969	Simulated	27,050	87,150	114,050	48,550	30,000	306,800
	Actual	40,322	95,610	130,398	66,256	35,383	369,647
1970	Simulated	29,050	92,000	124,200	53,250	33,500	332,000
	Actual	40,200	91,757	134,724	66,658	36,861	371,988
1971	Simulated	32,800	94,850	129,800	55,350	36,150	348,950
	Actual	41,307	89,210	130,395	64,630	34,852	362,187

\* Source: Vital Statistics - Statistics Canada Catalogue 84-201.





Table 7.3

Comparison of simulated and actual data \*

Marriages by Region: 1968-71

		Atlantic	Quebec	Ontario	Prairies	B.C.	Canada
1968	Simulated	18,200	54,050	70,775	35,525	20,850	199,400
	Actual	16,665	46,004	62,109	29,678	16,914	171,766
1969	Simulated	18,500	55,375	71,425	32,450	19,700	187,450
	Actual	17,420	47,545	67,150	31,378	18,284	182,183
1970	Simulated	19,800	55,600	73,625	34,900	19,950	203,875
	Actual	17,875	49,607	68,874	31,610	20,026	188,429
1971	Simulated	19,950	56,475	74,350	35,000	21,475	207,250
	Actual	18,678	49,695	69,590	32,554	20,389	191,324

Table 7.4

Comparison of simulated and actual data \*

Divorces by Region: 1968-71

		Atlantic	Quebec	Ontario	Prairies	B.C.	Canada
1968	Simulated	1,675	5,975	7,825	3,325	1,750	20,550
	Actual	675	606	5,036	2,765	2,220	11,343
1969	Simulated	1,550	6,075	7,800	3,550	1,825	20,800
	Actual	1,362	2,930	11,843	5,648	4,224	26,079
1970	Simulated	1,400	6,450	8,350	2,950	2,600	21,750
	Actual	1,414	4,865	12,451	5,876	5,111	29,775
1971	Simulated	2,000	5,975	8,450	2,800	2,400	21,625
	Actual	1,413	5,195	12,189	5,835	4,942	29,626

\* Source: Vital Statistics - Statistics Canada Catalogue 84-201.



married, and other. The "other" includes the widowed, separated and divorced. Including the separated into the eligible population for marriage clearly introduces positive biases.

It can be seen from table 7.5 that the death process results in an underestimate in the number of simulated deaths. This underestimate is 9.4%, 13.6%, 11.1%, and 8.2% respectively in the years 1968 through 1971. Most of this error (approximately 8%) is a consequence of the initial population underestimate, while simulation error can account for another +4%.

The effects of all of these various population flows are presented in Tables 7.6 and 7.7. Table 7.6 shows the magnitudes by which total population is changed by the flow processes of birth, death, immigration, and emigration. It also demonstrates the "Law of Conservation of Population". If the procedure by which new records are created through births and immigration are working properly, and if the procedure by which individual records are deleted through death and emigration are also working properly, then the final output population in any given year should equal the initial population plus the sum of births and immigrants less the sum of deaths and emigrants. The data in Table 7.6 indicates that the model does "conserve" population in this sense.



Table 7.5

Comparison of simulated and actual data\*

Deaths by Region: 1968-71

		Atlantic	Quebec	Ontario	Prairies	B.C.	Canada
1968	Simulated	14,750	33,200	50,800	24,350	15,800	138,900
	Actual	15,628	39,537	55,552	25,339	16,828	153,196
1969	Simulated	13,850	31,900	47,650	25,450	14,600	133,450
	Actual	15,524	40,103	55,707	25,453	17,377	154,477
1970	Simulated	13,650	33,300	51,000	23,900	15,200	137,050
	Actual	15,977	40,392	56,769	25,440	17,020	155,961
1971	Simulated	14,950	32,800	51,450	26,550	18,650	144,400
	Actual	15,831	40,738	56,623	25,963	17,783	157,272

\* Source: Vital Statistics - Statistics Canada Catalogue 84-201.



Table 7.6

Simulated population flows

	1967	1968	1969	1970
Initial Population	19,898,000	20,152,450	20,407,050	20,669,600
+				
Births	285,000	306,800	332,000	348,950
+				
Immigrants	180,250	157,850	144,200	118,450
-				
Deaths	138,900	133,450	137,050	144,400
-				
Emmigrants	71,900	76,600	76,600	75,850
=				
Predicted Simulated population at end of simulation year	20,152,450	20,407,050	20,669,600	20,916,750
Actual Simulated population at end of simulation year	20,152,450	20,407,050	20,669,600	20,916,750





Table 7.7 compares the sex-region populations produced by POLSIM with the same distributions as reported by Vital Statistics. It can be seen that there is a general underestimation on the part of the model. This can be explained by the fact that the Survey of Consumer Finance underestimates the total population (by excluding the Yukon and N.W.T., military personnel, and persons in institutions) and because the model itself underestimated the number of births in each of the simulated years. Over the four year period the underestimation in population that can be attributed to the model itself is 128,276 (sum of underestimates in births less sum of underestimates in deaths). The total underestimate in the 1971 simulated population is seen from table 7.7 to be 780,260. Of this total, 16% can be crudely attributed to the error generated by the model while 84% can be attributed to the error in the initial year population. This is in fact an upper estimate of the model error. The fact that the initial population is too small to begin with implies that one would expect an underestimate in the number of births and deaths. Therefore part of the underestimate arising from the model is in fact attributable to the error in initial year population. (A more sophisticated analysis of the error in the model itself can be carried out along the lines discussed in Chapter 4.)

The results of the labor force simulations are summarized briefly in table 7.8 and Figure 7.1. Table 7.8 presents the unemployment rates, by sex, that the model produces over the entire four year period and compares these



Table 7.7

Comparison of simulated and actual data \*

Population Distribution by Sex and Region: 1968-71

	Atlantic	Quebec	Ontario	Prairies	B.C.	Canada
Male Simulated	986,800	2,920,200	3,495,800	1,643,250	991,750	10,037,800
Male Actual	1,009,300	2,956,600	3,649,800	1,753,000	1,016,400	10,409,900
Female Simulated	977,450	2,952,600	3,541,850	1,615,400	953,150	10,040,450
Female Actual	991,700	2,970,400	3,656,200	1,704,000	990,600	10,334,100
Male Simulated	984,950	2,933,000	3,561,300	1,647,400	1,021,750	10,148,400
Male Actual	1,013,800	2,982,400	3,721,800	1,773,500	1,046,800	10,563,600
Female Simulated	973,100	2,962,700	3,619,850	1,621,250	985,000	10,161,900
Female Actual	998,200	3,001,600	3,730,200	1,725,500	1,020,200	10,497,400
Male Simulated	982,350	2,945,450	3,631,000	1,650,250	1,051,900	10,260,950
Male Actual	1,015,400	2,993,000	3,812,000	1,783,100	1,082,800	10,712,600
Female Simulated	972,250	2,974,800	3,696,200	1,628,500	1,018,550	10,290,300
Female Actual	1,002,600	3,020,000	3,825,000	1,739,900	1,054,200	10,664,400
Male Simulated	980,200	2,955,150	3,700,550	1,661,450	1,078,200	10,375,550
Male Actual	1,038,215	2,994,550	3,840,905	1,793,115	1,100,375	10,795,370
Female Simulated	973,100	2,983,750	3,767,850	1,641,500	1,046,300	10,412,500
Female Actual	1,019,040	3,033,215	3,862,200	1,749,240	1,084,245	10,772,940

\* Source: Vital Statistics - Statistics Canada Catalogue 84-201.

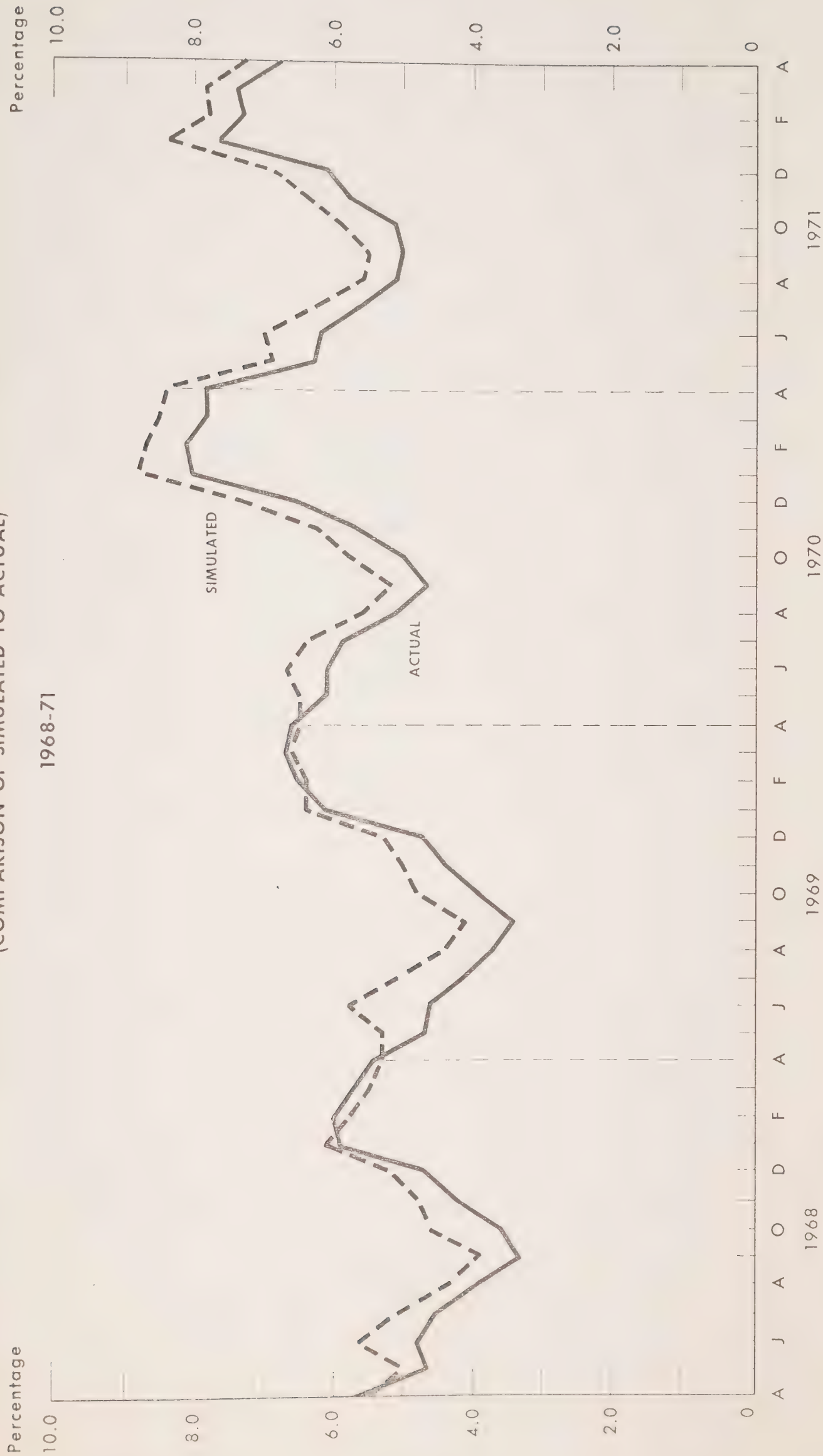


		APRIL	MAY	JUNE	JULY	AUG.	SEPT.	OCT.	NOV.	DEC.	JAN.	FEB.	MARCH	APRIL
Canada Total														
1968	Simulated	5.4	5.0	5.6	5.1	4.3	3.9	4.6	4.8	5.2	6.1	5.8	5.5	5.3
	Actual	5.7	4.6	4.8	4.5	3.9	3.3	3.6	4.2	4.7	5.9	6.0	5.7	5.4
1969	Simulated	5.3	5.3	5.8	5.1	4.4	4.1	4.8	5.0	5.3	6.4	6.4	6.6	6.5
	Actual	5.4	4.7	4.6	4.1	3.7	3.4	3.9	4.4	4.7	6.1	6.5	6.7	6.6
1970	Simulated	6.5	6.5	6.7	6.4	5.6	5.2	5.8	6.3	7.3	8.8	8.7	8.5	8.5
	Actual	6.6	6.1	6.1	5.9	5.1	4.7	5.0	5.7	6.5	8.0	8.1	7.8	7.8
1971	Simulated	8.4	6.9	7.0	6.3	5.6	5.5	5.9	6.4	6.9	8.4	7.8	7.9	7.3
	Actual	7.8	6.3	6.2	5.7	5.1	5.0	5.1	5.8	6.1	7.7	7.3	7.4	6.8
1968	Simulated	6.3	5.4	5.7	5.1	4.2	3.6	3.9	4.7	5.5	6.5	6.2	5.9	5.5
	Actual	6.7	5.3	5.1	4.7	3.9	3.3	3.8	4.8	5.4	6.8	7.1	6.7	6.2
1969	Simulated	5.5	5.6	5.8	5.2	4.3	4.0	4.4	5.0	5.7	7.1	7.1	7.4	7.1
	Actual	6.2	5.1	4.8	4.3	3.8	3.4	4.0	4.7	5.4	7.0	7.6	8.1	7.9
1970	Simulated	7.2	7.1	7.1	6.9	5.9	5.3	5.8	6.8	8.4	10.2	10.1	10.0	9.9
	Actual	7.9	6.9	6.5	6.2	5.3	5.0	5.2	6.1	7.3	9.2	9.3	9.2	9.0
1971	Simulated	9.7	7.7	7.5	6.8	5.8	5.6	5.9	6.9	7.8	9.9	9.1	9.4	8.3
	Actual	9.0	6.9	6.6	6.0	5.2	5.1	5.1	6.2	7.0	8.6	8.2	8.4	7.8
1968	Simulated	3.3	4.1	5.5	4.9	4.7	4.6	6.0	5.2	4.7	5.2	5.0	4.6	4.7
	Actual	3.4	3.3	4.2	3.9	3.7	3.2	3.3	3.1	3.1	4.0	3.7	3.4	3.5
1969	Simulated	4.8	4.9	5.7	4.9	4.7	4.5	5.6	4.9	4.5	5.1	5.2	4.9	5.1
	Actual	3.5	3.7	4.1	3.5	3.5	3.4	3.6	3.6	3.3	4.1	4.2	3.8	4.0
1970	Simulated	5.2	5.3	5.8	5.4	5.1	4.8	5.9	5.4	5.3	6.0	5.9	5.5	5.9
	Actual	4.0	4.2	5.2	5.3	4.7	4.3	4.6	4.7	4.7	5.5	5.5	5.0	5.5
1971	Simulated	6.0	5.4	6.0	5.2	5.0	5.2	6.0	5.5	5.1	5.7	5.4	5.1	5.3
	Actual	5.5	5.0	5.5	5.0	4.8	4.8	5.2	5.0	4.4	6.0	5.5	5.5	4.9



Figure 7.1

CANADA UNEMPLOYMENT RATES - TOTAL  
(COMPARISON OF SIMULATED TO ACTUAL)







with the actual rates for the same period as measured by the labor force survey. Figure 7.1 plots the monthly aggregate simulated unemployment rate and the monthly aggregate actual unemployment rate over the four year period. It can be seen that the model tracks the unemployment rate very well, and there is no tendency for it to get "off-track" as time progresses. There is a slight tendency, however, for the simulated rate to be too high. This was expected, because the simulation parameters had been adjusted to fit the higher unemployment rates of the period April 1972 to April 1973. (See Chapter 5). This adjustment was such as to increase the resulting simulated unemployment rate slightly from that which would have resulted from the original equations. Since the original regression equations of the labor-force model had been fitted to data from the 1959-1969 period, the adjustment would be expected to simulate too many unemployed persons over the four years 1967-71. The adjusted equations would be expected to perform better over a four year period beginning in 1971.

The Market Income simulation is summarized in Tables 7.9-7.11. For each of the component incomes (employment income, property income, and retirement income) distributions for the four simulated years are presented. As standards of comparison, the same distributions from the 1967 and 1971 SCF surveys are also given. It can be seen that the simulations perform as one would expect; there is a general tendency for the distributions to shift to the right as time progresses.



Table 7.9  
Employment Incomes

Income Categories in \$	Base '67	Final '68	Final '69	Final '70	Final '71	Base '71
- 999	13,480,300 66.90	12,590,950 62.48	12,454,450 61.03	12,522,250 60.58	12,533,000 59.92	13,933,050 64.76
K - 1999	842,250 4.18	1,272,500 6.31	1,258,800 6.17	1,211,950 5.86	1,152,900 5.51	799,550 3.72
K - 2999	879,700 4.37	1,253,400 6.22	1,264,500 6.20	1,233,950 5.97	1,227,050 5.87	687,700 3.20
K - 3999	998,550 4.95	1,012,100 5.02	1,070,200 5.24	1,082,550 5.24	1,095,550 5.24	771,400 3.59
K - 4999	999,100 4.96	900,000 4.47	957,800 4.69	970,800 4.70	933,800 4.46	820,400 3.81
K - 5999	917,900 4.55	752,350 3.73	756,650 3.71	769,000 3.72	834,850 3.99	774,150 3.60
K - 6999	711,300 3.53	638,950 3.17	595,300 2.92	610,800 2.96	629,200 3.01	705,500 3.28
- 7999	423,550 2.10	473,250 2.35	494,150 2.42	485,600 2.35	492,700 2.36	675,800 3.14
K - 8999	289,350 1.44	306,750 1.52	357,950 1.75	396,300 1.92	442,600 2.12	602,150 2.80
K - 9999	164,800 .82	240,900 1.20	257,950 1.26	259,600 1.26	271,250 1.30	451,250 2.10
K - 14999	300,850 1.49	559,650 2.78	747,650 3.66	874,700 4.23	957,850 4.58	953,500 4.43
K - 19999	72,900 .36	93,300 .46	132,050 .65	186,950 .90	249,300 1.19	189,400 .88
K - 24999	24,700 .12	24,850 .12	21,950 .11	12,600 .06	37,100 .18	57,950 .27
K +	47,200 .23	33,500 .17	37,650 .18	52,550 .25	59,600 .28	93,850 .44
total	20,152,450 99.99	20,152,450 100.00	20,407,050 99.99	20,669,600 100.00	20,916,750 100.01	21,515,650 100.02



Table 7.10  
Property Incomes

Income Categories in \$	Base '67	Final '68	Final '69	Final '70	Final '71	Base '71
0	18,417,450 91.39%	16,597,100 82.36	15,666,000 76.77	15,119,550 73.15	14,806,200 70.79	18,008,550 83.70
1 - 250	857,900 4.26	2,525,700 12.53	3,460,350 16.96	4,010,650 19.40	4,331,250 20.71	2,200,400 10.23
51 - 500	268,200 1.33	298,350 1.48	405,650 1.99	503,100 2.43	589,850 2.82	413,450 1.92
01 - 750	152,600 .76	187,400 .93	224,050 1.10	271,850 1.32	306,100 1.46	202,500 .94
51 - 1000	117,000 .58	115,700 .57	142,600 .70	172,400 .83	196,350 .94	157,850 .73
1K - 1999	183,650 .91	235,300 1.17	267,900 1.31	308,300 1.49	353,800 1.69	268,400 1.25
2K - 2999	72,500 .36	74,650 .37	89,350 .44	103,950 .50	122,050 .58	111,350 .52
3K - 3999	29,600 .15	51,200 .25	64,200 .31	74,100 .36	83,900 .40	54,350 .25
4K - 4999	16,900 .08	22,050 .11	26,900 .13	33,900 .16	40,400 .19	34,500 .16
5K - 5999	10,850 .05	13,700 .07	19,200 .09	21,000 .10	24,950 .12	15,300 .07
6K - 6999	7,800 .04	7,350 .04	9,050 .04	11,300 .05	14,500 .07	13,650 .06
7K - 7999	4,850 .02	7,650 .04	9,200 .05	10,100 .05	12,450 .06	7,250 .03
8K +	13,150 .07	16,300 .08	22,600 .11	29,400 .14	34,950 .17	28,100 .13
total	20,152,450 100.00	20,152,450 100.00	20,407,050 100.00	20,669,600 99.98	20,916,750 100.00	21,515,650 99.99



Table 7.11  
Retirement Incomes

Income Categories in \$	Base '67	Final '68	Final '69	Final '70	Final '71	Base '71
0	19,775,350 98.13%	19,768,300 98.09	20,002,200 98.02	20,244,100 97.94	20,474,100 97.88	20,989,150 97.55
1 - 250	41,650 .21	43,400 .22	49,300 .24	53,800 .26	57,050 .27	62,350 .29
1 - 500	51,200 .25	50,850 .25	53,850 .26	56,100 .27	57,200 .27	66,700 .31
1 - 750	41,750 .21	40,600 .20	42,850 .21	44,400 .21	46,300 .22	51,800 .24
1 - 1000	45,250 .22	46,550 .23	46,600 .23	47,550 .23	48,100 .23	47,700 .22
K - 1999	102,500 .51	104,250 .52	107,100 .52	109,700 .53	112,000 .54	127,100 .59
K - 2999	52,650 .26	54,250 .27	55,800 .27	58,750 .28	59,900 .29	73,450 .34
K - 3999	22,700 .11	23,350 .11	24,250 .12	26,000 .13	27,800 .13	42,200 .20
K - 4999	11,700 .06	10,550 .05	11,750 .06	13,500 .07	15,750 .08	26,300 .12
K - 5999	4,200 .02	5,150 .03	6,200 .03	6,600 .03	7,800 .04	11,850 .06
K - 6999	1,400 .01	2,250 .01	2,850 .01	3,750 .02	4,400 .02	6,400 .03
K - 7999	1,050 .01	1,150 .01	1,600 .01	1,800 .01	1,850 .01	3,100 .01
K +	1,050 .01	1,800 .01	2,700 .01	3,550 .02	4,500 .02	7,550 .04
tal	20,152,450 100.01	20,152,450 100.00	20,407,050 99.99	20,669,600 100.00	20,916,750 100.00	21,515,650 100.00





Since no adequate income data exists for the years between 1967 and 1971, it is not possible to examine how well the simulations perform year by year. All that can be compared are the final results (the distributions for the year 1971). It should be noted that the comparisons for property income in Table 7.10 are not really meaningful, due to the fact that the 1967 survey did not distinguish between dividend and interest income. As a result, the model simulated total property income with interest income transition matrices, and hence the resulting final simulated property income is not strictly comparable with the results obtained from the 1971 SCF survey.

A more detailed comparison between the final 1971 simulated results and the actual data for 1971 are given in the next section.

### 7.3 Analysis of the Simulated 1971 Population

The comparisons between the 1971 simulated distributions and the 1971 distributions as measured by the SCF survey are presented in Tables 7.12-7.24. These tables show how the populations in the two samples are distributed over demographic characteristics (age, sex, province, etc.), activity variables (weeks employed etc.), and market income characteristics.



Table 7.12

Comparison of simulated and base year populations by province for 1971

Province	Simulated	Base Year
Newf	487,550 2.33%	523,300 2.46%
PEI	107,250 0.51%	107,700 0.51%
NS	768,650 3.67%	795,450 3.74%
NB	586,250 2.80%	622,850 2.93%
PO	5,945,400 28.42%	5,996,950 28.16%
Ont	7,553,050 36.11%	7,645,750 35.91%
Man	960,450 4.59%	963,450 4.52%
Sask	830,800 3.97%	850,450 3.99%
Alta	1,510,000 7.22%	1,600,900 7.52%
BC	2,167,350 10.36%	2,186,400 10.27%
TOTAL	20,916,750 100.00%	21,293,200 100.00%



Table 7.13

Comparison of simulated and base year populations by region for 1971

Region	Simulated	Base Year
Maritime	1,949,700 9.32%	2,049,300 9.62%
PQ	5,945,400 28.42%	5,996,950 28.16%
Ont	7,553,050 36.11%	7,645,750 35.91%
Prairies	3,301,250 15.78%	3,414,800 16.04%
BC	2,167,350 10.36%	2,186,400 10.27%
TOTAL	20,916,750 100.00%	21,293,200 100.00%



Table 7.14

Comparison of simulated and base year populations by age groups  
for 1971

Age	Simulated	Base Year
0-14	6,148,150 29.39%	6,309,300 29.63%
15-19	2,098,350 10.03%	2,074,600 9.74%
20-24	1,727,450 8.26%	1,818,200 8.54%
25-29	1,612,600 7.71%	1,589,100 7.46%
30-34	1,243,000 5.94%	1,342,800 6.31%
35-39	1,280,300 6.12%	1,257,450 5.91%
40-44	1,237,500 5.92%	1,290,500 6.06%
45-49	1,159,950 5.55%	1,217,150 5.72%
50-54	1,030,350 4.93%	1,054,250 4.95%
55-59	908,750 4.34%	885,800 4.16%
60-64	767,650 3.67%	740,950 3.48%
65+	1,702,700 8.14%	1,713,100 8.05%
TOTAL	20,916,750 100.00%	21,293,200 100.00%





Table 7.15

Comparison of simulated and base year populations by sex  
for 1971

Sex	Simulated	Base Year
Male	10,431,400 49.86%	10,642,300 49.98%
Female	10,485,350 50.13%	10,650,400 50.02%
TOTAL	20,916,750 100.00%	21,293,200 100.00%

Table 7.16

Comparison of simulated and base year populations by family status  
for 1971

Family Status	Simulated	Base Year
Unattached	2,210,650 10.57%	2,405,100 11.3%
Family Head	5,212,850 24.92%	5,146,050 24.17%
Wife	4,753,250 22.72%	4,702,850 22.09%
Children	8,740,000 41.79%	9,039,200 42.45%
TOTAL	20,916,750 100.00%	21,293,200 100.00%



Table 7.17

Comparison of simulated and base year populations by marital status  
for 1971

Marital Status	Simulated	Base Year
Single	10,097,450 48.27%	10,435,050 49.01%
Married	9,544,800 45.63%	9,462,600 44.44%
Other	1,274,500 6.09%	1,395,550 6.55%
TOTAL	20,916,750 100.00%	21,293,200 100.00%

Table 7.18

Comparison of simulated and base year populations by number of weeks in school  
for 1971

Weeks in School	Simulated	Base Year
0	14,680,850 70.19%	13,639,300 64.06%
1-12	317,000 1.52%	11,950 0.06%
13-28	620,400 2.97%	344,500 1.62%
29-44	5,298,500 25.33%	7,297,450 34.90%
45+	0 0%	0 0%
TOTAL	20,916,750 100.00%	21,293,200 100.00%



Table 7.19

Comparison of simulated and base year populations by number of weeks employed  
for 1971

Number of Weeks Employed	Simulated	Base Year
0	11,439,550 54.69%	12,004,100 56.38%
1-12	590,200 2.82%	853,100 4.01%
13-24	1,202,500 5.75%	655,100 3.08%
25-36	1,182,700 5.66%	763,900 3.59%
37-48	1,810,550 8.66%	712,050 3.34%
49-52	4,691,250 22.43%	6,304,650 29.61%
TOTAL	20,916,750 100.00%	21,293,200 100.00%



Table 7.20

Comparison of simulated and base year populations by number of weeks unemployed  
for 1971

Number of Weeks Unemployed	Simulated	Base Year
0	18,773,550 89.76%	19,635,250 92.22%
1-12	1,176,450 5.62%	709,450 3.34%
13-24	663,100 3.17%	377,650 1.77%
25-36	215,200 1.03%	293,750 1.39%
37-48	73,200 0.34%	185,400 0.87%
49-52	15,250 0.07%	91,700 0.43%
TOTAL	20,916,750 100.00%	21,293,200 100.00%





Table 7.21

Comparison of simulated and base year populations by number of weeks  
in non-labour force for 1971

Number of Weeks in Non-Labour Force	Simulated	Base Year
0	6,574,350 31.43%	7,473,400 35.10%
1-12	5,343,900 25.55%	7,815,450 36.71%
13-24	1,256,600 6.01%	501,500 2.36%
25-36	775,950 3.70%	426,400 2.00%
37-48	986,700 4.72%	353,550 1.66%
49-52	5,979,250 28.59%	4,722,900 22.18%
TOTAL	20,916,750 100.00%	21,293,200 100.00%



Table 7.22

Comparison of simulated and base year populations by employment income  
categories for 1971

Employment Income	Simulated	Base Year
No Income	11,471,050 54.84%	12,054,750 56.61%
\$1-499	437,800 2.09%	896,550 4.21%
\$500-999	624,150 2.98%	608,200 2.86%
\$1K-1499	607,600 2.90%	458,650 2.15%
\$1500-2K	545,300 2.61%	389,200 1.83%
\$2K-2499	611,000 2.92%	368,300 1.73%
\$2500-3K	616,050 2.95%	332,950 1.56%
\$3K-3999	1,095,550 5.24%	788,650 3.7%
\$4K-4999	933,800 4.46%	798,450 3.75%
\$5K-5999	834,850 3.99%	791,600 3.72%
\$6K-6999	629,200 3.01%	725,550 3.41%
\$7K-7999	492,700 2.36%	689,750 3.24%
\$8K-9999	713,850 3.41%	1,063,200 4.99%
\$10K-12K	448,900 2.15%	590,300 2.77%
\$12K-15K	508,950 2.43%	403,700 1.90%
\$15K-25K	286,400 1.37%	256,250 1.20%
\$25,000+	59,600 0.28%	77,150 0.36%
TOTAL	20,916,750 100.00%	21,293,200 100.00%



Table 7.23

Comparison of simulated and base year populations by interest  
income categories for 1971

Interest Income	Simulated	Base Year
No Income	14,806,200 70.79%	17,723,200 83.24%
\$1-249	4,331,250 20.71%	2,217,450 10.41%
\$250-499	589,850 2.82%	416,550 1.96%
\$500-749	306,100 1.46%	224,700 1.06%
\$749-999	196,350 0.94%	151,450 0.71%
\$1K-2K	353,800 1.69%	281,200 1.32%
\$2K-3K	122,050 0.58%	114,950 0.54%
\$3K-4K	83,900 0.40%	59,550 0.28%
\$4K-5K	40,400 0.19%	36,000 0.17%
\$5K-8K	51,900 0.25%	38,650 0.18%
\$8K+	34,950 0.17%	29,500 0.14%
TOTAL	20,916,750 100.00%	21,293,200 100.00%



Table 7.24

Comparison of simulated and base year populations by retirement income categories for 1971

Retirement Income	Simulated	Base Year
No Income	20,474,100 97.89%	20,744,400 97.43%
\$1-249	56,550 0.27%	63,550 0.3%
\$250-499	57,100 0.27%	68,050 0.32%
\$500-749	46,350 0.22%	55,600 0.26%
\$750-999	47,400 0.23%	49,300 0.23%
\$1K-1499	64,650 0.31%	72,250 0.34%
\$1500-2K	47,850 0.23%	61,400 0.29%
\$2K-3K	59,800 0.29%	75,950 0.36%
\$3K-4K	28,050 0.13%	45,250 0.21%
\$4K-5K	15,550 0.07%	25,000 0.12%
\$5K-8K	14,850 0.07%	24,200 0.11%
\$8K+	4,500 0.02%	8,250 0.04%
TOTAL	20,916,750 100.00%	21,293,200 100.00%

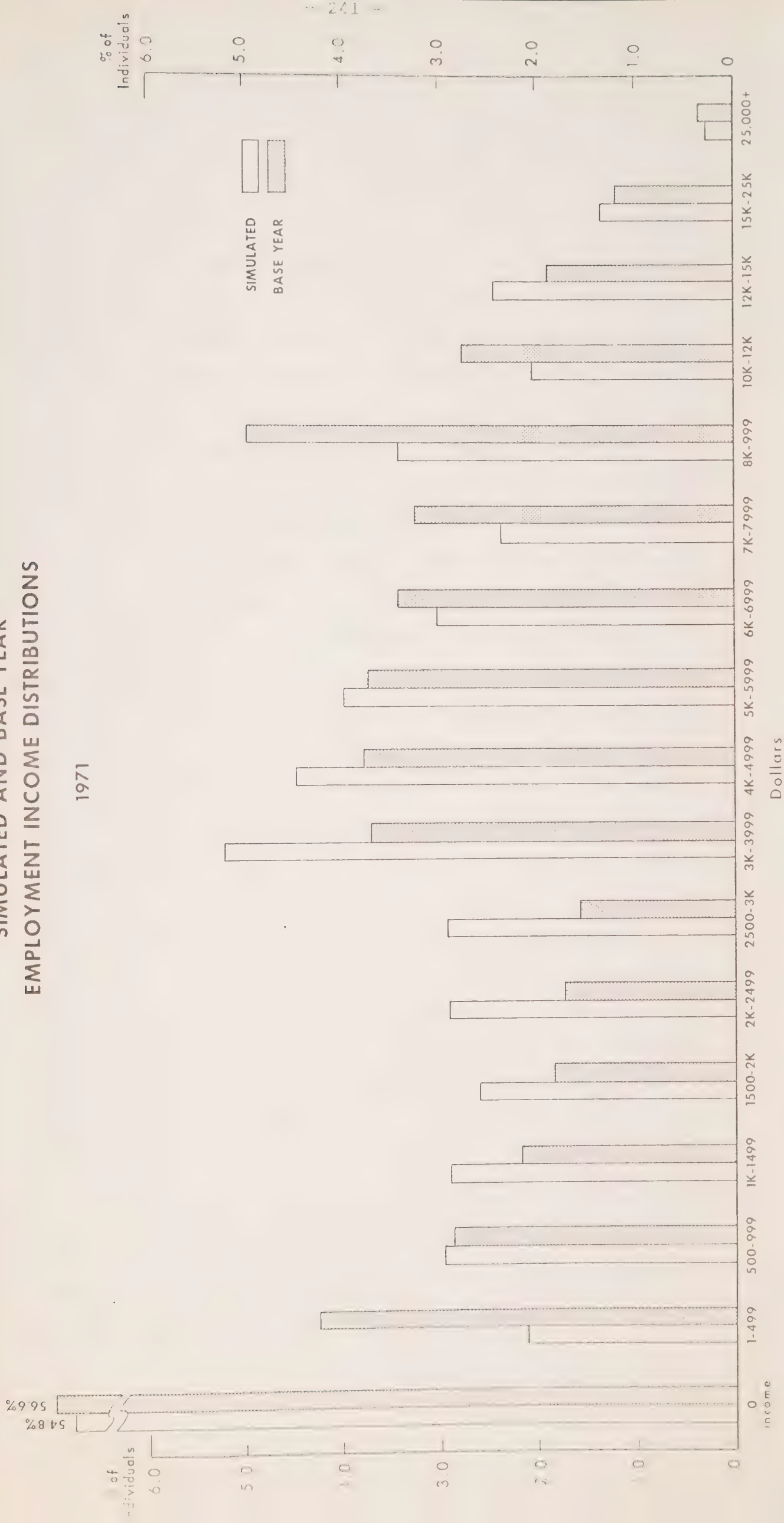




Figure 7.2

# SIMULATED AND BASE YEAR EMPLOYMENT INCOME DISTRIBUTIONS

1971





In assessing these distributions it is not possible to assert that they are "good" or "bad" in any objective sense. What we have to do is come up with some notion of whether the simulated population, when viewed comprehensively (over all pertinent variables), is adequate for the purposes to which it is to be put. This is clearly not a question which permits an unequivocal answer. Whether or not the simulation is "adequate" will depend on a number of factors. It will depend on the particular purpose for which the simulation is being used, on alternative sources of data, on the level of dissagregation at which we wish to view the results, and so on. And in the final analysis, it will also depend on the user's subjective idea of just what "adequate" means for his particular objective. Rather than attempting to answer the question of adequacy once and for all, the present statement will simply report the results that were obtained. It will be left to the individual user to determine how "good" these results are.

Tables 7.12-7.17 summarize the results of the Demographic Block simulation. By the standard of comparative final distributions, the demographic block is the most satisfactory part of the POLSIM model. It can be seen that over all of the demographic variables the 1971 actual and 1971 simulated distributions are very close indeed.

This is to be expected, of course, since demographic characteristics are either easy to simulate (age and sex for example) or else affect small proportions of the population (death, emigration, etc.). They would therefore be expected



to compare closely to the base year data. There is a discrepancy in the total populations, however, which should be noted. The simulated final population is 20,916,750 while the 1971 survey estimated population is 21,292,640. The 1971 survey estimated population is therefore 1.79% higher than was obtained through the simulation. The reason for this discrepancy is complex, since it depends on the simulation of births, deaths, immigrants, and emigrants, as well as on the system of weighting that is used by the SCF survey. The errors in the simulation proper have already been discussed. The immigration totals accounted for a cumulative underestimate of 14,368 persons. It is difficult to say anything valid at all about the emigration estimates, because there exists no accurate data on the true extent of emigration. For the present we will ignore any underestimate that might occur on this account. Births and deaths result in a cumulative underestimate of 127,256 persons. The total underestimate resulting from the simulation is therefore approximately 141,624 persons, which means that it is still necessary to account for an underestimate of approximately 234,266 persons.

This can be explained by the way in which the SCF surveys are weighted. The weights for the 1967 SCF apply to the population as of December 1967, while the weights for the 1971 SCF apply to the population as of June 1972. This is a difference of four and one half years. The simulation, on the other hand, is for four years only, and hence one would expect the 1971 SCF population to slightly exceed the population generated by the simulation.



The Activity characteristics, which are summarized in Tables 7.18-7.21, would not be expected to fare so well as the demographic variables. This is because the activity variables are simulated indirectly (through a month to month simulation), and because the whole population is subject to extensive changes. The Tables indicate that on the whole the simulated activity characteristics, though not as close as are the demographic characteristics, are still fairly close to the base year distributions.

The most important of the activity variables is weeks of unemployment, and this variable is given in Table 7.20. It can be seen that the simulated distribution is more "flat" than the base year distribution; the simulation tends to simulate unemployment for a relatively larger group of people, but at the same to make the duration of this unemployment shorter. Thus the simulation "unemploys" 10.24% of the population for at least some period, compared with 7.78% in the base year. In the simulation, however, 54.8% of this group is unemployed for less than 12 weeks, compared with 42.9% for the base year group. Interestingly, the total number of weeks of unemployment over all individuals is only slightly less for the simulated group (28.1 million weeks versus 29.1 million for the base year). This "flattening" of the distribution is a consequence of the Markov-one nature of the simulation. In a Markov-one process, a person's present state is assumed to depend only on his state in the immediately preceeding period. In this instance, this is clearly not an adequate assumption. Whether or not a person is to become unemployed depends on his employment history for several preceeding periods. This is a refinement that may be dealt with in future versions of the model.





The results of the Market Income simulation are presented in Tables 7.22-7.24. The tables do not include the results of a simulation of dividend income due to the fact that dividends were not distinguished from other property income in the 1967 survey. As mentioned earlier, the simulation moved all of this property income forward as if it were interest income. The interest income comparisons can therefore not be expected to be very close. They do indicate, however, that the interest income transition matrices will not produce incomes that are wildly off-track.

The results of the employment income simulation may be examined in figure 7.2 and table 7.22. It can be seen that the simulated distribution is too low at the low end of the income scale, becomes too high for the lower-middle income groups, becomes too low again for the upper-middle income classes, and then finally is too high for the higher income groups. The tendency for the simulated distribution to exhibit a lower variance than the base year distribution, at least among the low and middle income groups, may be explained in part by the simulation of unemployment. As explained above, the Markov-one nature of the unemployment simulation tends to flatten the distribution of weeks of unemployment. That is, too many people have small amounts of simulated unemployment while too few have large amounts of simulated unemployment. This bias in the unemployment simulation will affect Class B persons (approximately 80% of the labor force - those who are in occupations which are likely to result in at least some unemployment) but not Class A persons (the remaining persons in the labor force, the 20% whose occupations are such that they will never likely



become unemployed). The flattening of the unemployment distribution among Class B persons will tend to lower the variance in employment incomes. There will be too few simulated persons with very low incomes (which are a result of high unemployment), and too many with low to middle incomes (a result of too many relatively low wage persons - the Class B group - with at least some unemployment experience). This effect will be mitigated as income increases because Class A persons will increasingly tend to dominate the distribution as incomes increase.

It will be noted that the employment income simulation is not as good as would be expected from a cursory examination of the results of Chapter 6. Figures 6.2b and 6.2f present the results of a validation of the market income block parameters for prime age males, the group which forms the largest proportion of persons in the labor force. These graphs demonstrate that the annual wage transition matrices and weekly wage transition matrices are such as to reproduce annual and weekly wage distributions over a four year period almost exactly. These expected 1971 distributions do not include any simulation error, since they are produced by multiplying the 1967 distributions by the fourth power of the relevant matrix. But one would expect a very small simulation error in any case, due to the large number of persons in the simulation. So an examination of Chapter 6 alone would lead one to expect an excellent simulation of employment income. But for the reasons stated above, the actual simulation does not yield the almost perfect results that might otherwise be expected.



A better simulation of employment income probably could be produced, provided one wanted to simulate income independently of the underlying labor force activity that produces employment income. It is the attempt to explicitly model labour force activity which creates difficulties in the present model. We have indicated above ways in which these difficulties may be overcome in future versions of the model.

The retirement income simulation, as can be seen from Table 7.24, is very good. It tends to follow the base year distribution almost exactly.



## 8. The Policy Block

In one sense all of the other blocks of POLSIM are a prelude to the Policy Block. The ultimate objective of advancing the model population through time is not achieved until the Policy Block, with its models of government programs, has been run. In this chapter we shall first consider the problem of evaluating government programs from a purely technical point of view in the context of POLSIM. We shall then briefly consider the program models or policy algorithms which have been developed as part of the POLSIM project, leaving a fuller discussion of these models to a later report. Next we shall indicate how the Policy Block is run. Finally, we give an example of the application of a policy algorithm.

### 8.1 Evaluating the Effects of Government Programs

In general there are two classes of effects caused by government programs. The first we call real effects (e.g. changes in relative prices, changes in work effort, etc.) and the second we refer to as financial effects (e.g. changes in disposable money income). In modelling the operation of government programs, our microcomponents are cast in a particular macroeconomic environment. This macroeconomic environment is defined in the model by the exogenous specification of such things as price indices and unemployment rates, which then translate into real effects (e.g. unemployment) for particular individuals. These real effects are determined before the Policy Block begins. That is, the real effects do not depend in any formal way on the government programs contained in the Policy Block. Neither do the policy algorithms





of the Policy Block produce real effects directly on the microcomponents. Individual behavior is unaffected by the particular policies modelled. Financial effects, then, are calculated under the assumption that there is no feedback from the policy algorithms to real effects.

This kind of treatment is rigid if not unrealistic, and it does limit the usefulness of the model. However, it does not mean that no recognition whatever can be made of probable behavioral effects caused by individual government programs. For instance, in the case of tax-transfer work disincentive effects it is possible to build behavioral response into the model. Nevertheless, the existing structure of the model does mean that a given time track will remain undisturbed by alterations of policy algorithms in the Policy Block. That is, possible current behavioral effects of individual government programs, while they can be made to affect today's outcomes, do not affect tomorrow's possibilities or events. This we regard as the most serious limitation of the POLSIM model and it arises mainly because of the absence of explicit treatment of capital stocks in the model.

## 8.2 Policy Algorithms

A number of policy algorithms have been constructed or are under development as part of the POLSIM project. For the moment we shall do no more than list the names of the government programs modelled at this time. A later report will document software and contain tests performed to establish the accuracy of simulation results achieved using these algorithms. Algorithms exist for the following government programs:



- (a) Personal Income Taxes (federal and provincial)
- (b) Old Age Security
- (c) Guaranteed Income Supplement
- (d) Canada (Quebec) Pension Plan Contributions
- (e) Unemployment Insurance Premia
- (f) Hypothetical Negative Income Tax Programs

### 8.3 Running the Policy Block

The Policy Block, comprised as it is of a series of individual algorithms, does not possess a structure which is in any way similar to the other blocks. The Policy Block proper is embodied in a computer program, RESULT, whose function is to call subroutines expressing the policy algorithms and to accumulate and print out in convenient tabular form the effects of these policies. Program RESULT (see Appendix F) can be readily altered to accommodate a wide variety of reporting formats.

The Policy Block takes as input the file describing the model population for some given year and proceeds to produce the program effects for that same year. Since the Policy Block does not affect any given time track, it may be run either for all years of a given projection or for selected years only.



The computer program used for producing the distribution tables cross-classifies the policy effects in a number of different ways. There are twenty-two different classifications (e.g., province, sex, marital status, total income, etc.). Each classification is divided into a varying number of categories (e.g. ten categories for the provinces, two for sex, seventeen for total income, etc.). Any combination of pairs of different classifications can be used to produce the desired cross-classification tables, up to a maximum of 22 cross-classifications.

The program will read either an individual's record or a family's records, depending on which input is required for the policy being studied. The user specifies his choice by setting the value of an input flag. An output flag must also be set by the user. This flag determines whether distributions of individuals are required in the output tables or whether distributions of families are required. Family distributions are produced from the characteristics of the head of the family in all cases but income. The sum of the family's incomes is used for assignment to income categories of family distributions.

After reading each record (or records in the case of families being read) the program determines which cells in the output tables are relevant for that particular record. Each characteristic (e.g. province, sex, income) of the individual or family head is assigned to the appropriate category for that characteristic. (For example, the province category for an individual from Newfoundland is one, from B.C. it is ten; the income category for no income is one, for a total income above \$25,000 it is 17). In the case of a



family's records being read and individual distributions produced, each member of the family is treated as an individual, and each characteristic of each family member is assigned to the appropriate category.

Once it has been determined where in the pre-specified distributional categories the given individual (or family) will fall, the micro-effects of the program to be studied are calculated for this particular unit. This requires calling the appropriate policy algorithm which computes the desired effects. A tax algorithm, for example, would calculate the total taxes paid by a given family. A UIC contribution algorithm would calculate the UIC contributions payable by a given individual.

The effects thus determined are then added to the effects calculated for all other persons in the same category. For example, if the person is from Newfoundland and has a total income of \$7,000, the taxes he pays would be added to the taxes paid by all other individuals from Newfoundland in the \$7,000-\$8,000 income bracket.

The program proceeds in this fashion until all individuals or families have been read. It then prints out the cross-classification tables desired by the user. These tables present both the absolute effects and the percentage distributional effects. The program listing and an example of the program output is given in Appendix F.





#### 8.4 Example of Policy Simulation: The Personal Income Tax Algorithm

The Personal Income Tax algorithm is a computer program designed to compute the effects of the personal income tax on the family records that are output of the POLSIM model. The input to the program is the POLSIM state vector of a single family, the CPP and UIC contributions of each individual family member (these quantities it should be noted are themselves outputs of policy algorithms), the inflation factor necessary to adjust tax brackets and exemption levels, and the year for which the simulation is to apply. The output, for each family member, is the following: his income (taxation definition), his "tax status" (whether he is an unmarried family head, a dependent, a married man whose wife is deductible, a married woman whose husband is deductible, a married person who has a larger income than his spouse who files a separate return, a married person who has a smaller income than his spouse who files a separate return, or an independent child), his federal taxable income, his federal tax payable, his provincial tax payable, and if he is a resident of Quebec, his Quebec taxable income.

The program in its present version incorporates all of the changes in the tax legislation up to and including the February 19, 1973 Budget (see Appendix F). It begins by calculating the basic tax parameters for the given year: the various exemption levels as determined by the inflation index, and the marginal tax rate to apply in the first tax bracket for the year being simulated. The program then calculates each family member's taxation income. Employment



expenses, tuition, and CPP-UIC contributions are deducted if applicable, and dividend income is grossed up by 1/3. The person's income includes those items present on the individual state vector (employment income, interest, dividends, retirement income, and other money income) plus income items which are produced by other policy algorithms and are taxable. No attempt is made in the present version of the model to impute capital gains.

Once the individual incomes are calculated, tax status is determined. This then enables all of the various applicable deductions to be calculated for each family member: the basic exemption for an adult, the standard charity-medical deduction (assumed to be \$100.00 for all persons), the old-age deduction, the marriage equivalent deduction for a dependent in the absence of a spouse, the deductions for children (which depend on the child's income), and the spouse deductions. Taxable income is thus determined (income minus total deductions), and the program proceeds to the calculation of tax. This is done by determining which tax bracket the person is in, and then summing the tax payable at the beginning of that bracket with the marginal tax payable on the income within the given bracket. Provincial tax is then calculated, and the program returns to receive the record of another family.

It is not possible to completely validate the present tax algorithm. To do so would require that we simulate 1973 taxes and then compare our results with data compiled by DNR for the 1973 taxation year. And this in turn requires



that DNR data be available. Unfortunately, the most recent DNR data that exists is for the 1971 taxation year, and this is also the year of the most recent SCF survey. That is, it is the most recent year for which a base year population, as opposed to a simulated population, is available for policy simulations.

These data limitations resulted in something of a dilemma, insofar as validating the tax simulation was concerned, because we did not possess a 1971 tax algorithm. Indeed, there is no reason why we would construct one, because the purpose of POLSIM is to simulate policies into the future, not into the past. As a compromise it was decided to use the 1972 tax algorithm (an earlier version of the algorithm described above), to simulate 1971 taxes. This would not give us an ideal check on the validity of the algorithm, but it would enable us to check whether the algorithm produces severe distortions. To the extent that the changes introduced by the 1972 tax reform did not grossly change either the absolute rate of tax (on total assessed income), or the distribution of tax across income categories, the simulation could be expected to compare very well with the data collected by DNR.

The results of the simulation are summarized in Table 8.1. Two simulations were carried out, one on the 1971 SCF population, and the other on the 1971 population generated from the 1967 base year by the POLSIM model. Both of these simulations are then compared with the data derived from DNR records.



Table 8.1

SIMULATION OF TOTAL TAX BY INCOME CLASS

	1971 DNR Statistics DOLLARS (000's)	Simulation From 1971 SCF Population* DOLLARS (000's)	Simulation From 1971 POLSIM Population* DOLLARS (000's)
-1499	2,526 0.03%	0 0.00%	0 0.00%
500-2499	58,055 0.70%	31,510 0.36%	56,142 0.67%
500-3999	347,596 4.17%	258,906 2.92%	402,334 4.78%
4K-4999	422,448 5.07%	355,700 4.01%	431,404 5.12%
5K-5999	531,912 6.38%	494,530 5.57%	535,072 6.35%
6K-6999	618,218 7.42%	616,905 6.95%	532,744 6.32%
7K-7999	720,622 8.65%	732,240 8.25%	521,807 6.19%
8K-9999	1,394,416 16.74%	1,532,760 17.28%	1,005,235 11.93%
\$0K-12K	1,037,435 12.45%	1,184,475 13.35%	922,128 10.95%
\$2K-15K	917,325 11.01%	1,096,040 12.36%	1,399,999 16.62%
\$5K+	2,280,333 27.37%	2,567,929 28.94%	2,616,379 31.06%
TOTAL	8,330,886 100.00%	8,870,995 100.00%	8,423,244 100.00%

Simulations of total tax using the 1972 tax structure





It can be seen that the simulation performs very well. The simulation on the 1971 SCF population matches the actual data very closely, both in terms of the distribution and in the absolute amounts. The simulation does, however, tend to slightly underestimate tax at the low end of the income scale and to overestimate it at the upper end. Both of these tendencies are what one would expect. To begin with, one of the objectives of the 1971 tax reform was "to give tax relief to Canadians of lower incomes". Thus the application of the 1972 rules to the 1971 population would in itself be sufficient to shift the distribution to the upper end of the income scale. In addition to this, there are other reasons why the simulation would tend to shift the distribution of tax. In actual practice it is possible for persons with very low assessed income to pay relatively large amounts of tax. This situation can arise because: (1) returns may be filed by non-residents of Canada in respect of income from Canada which is not subject to personal exemptions; (2) individuals who are resident in Canada for only part of a taxation year will have their exemptions pro-rated to the period in which they earned their income; and (3) some returns are taxable only in respect of lump sum pension payments which are excluded from total income. It is not possible at present to simulate any of these effects. In addition to these effects on the low income classes, simulated taxes in the higher income classes will tend to be overestimated. This is because the simulation only takes account of the minimum exemptions that could arise. Other exemptions which are not accounted for could arise because business or farm losses of earlier years may offset the current year's income, or because of such factors as foreign tax credits, registered retirement plans, unusual medical expenses, allowable deductions from investment income, or gifts to the Crown.



The tax simulation on the 1971 simulated population can be seen to compare quite closely with the simulation on the base year population, at least in terms of total taxes generated. The distributions are different, however, which is a result of the differences in the distribution of income. The extent and reasons for these differences in the income distributions are documented in Chapter 7.

The above problems notwithstanding, it is possible to conclude from the present validation that the existing tax algorithm performs more than adequately. A more conclusive judgement will become possible when the taxation statistics for the 1972 taxation year are released by DNR.



## 9. Concluding Remarks

This chapter is concerned to set down some of the most important things we have learned about micro-data modelling in general and to suggest improvements which may be made in the POLSIM model in particular.

The first mentioned can be disposed of fairly quickly. These may be summarized in three statements. First, micro-data modelling of whole populations is quite expensive, particularly in the model development stage. Our problems were perhaps exacerbated because of the necessity of utilizing confidential data at Statistics Canada but the fact remains that one is manipulating large amounts of data and this can be costly both in terms of time and money. Cost can be lowered, of course, by utilizing more efficient computing systems and by smaller samples in certain instances. We have done some work on the former question but not on the latter. It may be true, for example, that most of the processes we were concerned to model can be done well enough with a sample half the size of the one utilized.

Second, it is important to commence information interchange between model analysts and computer systems experts at a very early stage of model development. This also relates to the question of cost. There is no reason why the most efficient programming of the model cannot be the first programming. The chances of achieving this are obviously immeasurably greater if this question is raised at the time of the development of the model structure.



Third, it is to be preferred if the applications of the microdata model can be well defined early on. This would mean that the absolute minimum of data necessary for the individual state vector could be quickly identified and costly changes avoided. The composition of the individual state vector adopted for the present version of POLSIM represents a notional compromise between policy issues of assumed importance, adequate detail for modelling in relation to these policy issues, and cost. This process of compromise will be easier in situations where the policy questions can be well specified in advance.

We now turn to the question of particular improvements to POLSIM. One feature, not of POLSIM proper but of the simulation exercise, which should receive attention is the adjustment of the initial year sample to better align it with other population measures. Given enough independent data for this purpose, it should be possible to bring initial population errors close to zero.

In the Demographic Block there is an obvious need to make certain of the probability parameters, for example the fertility probabilities, time variant. There is also a need to extend the stratifications on some of these variables. The divorce probabilities, for example, could be stratified by the characteristics of both spouses, rather than just one. And marriage probabilities could be made conditional on the education of the spouse. Theoretical analysis can also be easily extended in the Demographic Block. For example, one could develop a more rigorous model of the divorce process. Is divorce to be restricted to legal separation or should it be extended to include any type of separation?





All of the parameters in the Immigration Block are time invariant and derived from the data available for the 1971 year. In particular, the province-age-sex-marital status distribution of new immigrants applies to the year 1971 alone. It would clearly be desirable to estimate this distribution, from time series data, as a function of (say) economic conditions in Canada and abroad and possibly other variables as well. It would also be useful to attempt a model which predicted the total number of immigrants as a function of economic conditions, both in Canada and in the largest countries from which immigrants come.

Several extensions are possible for the Activity Block:

- (i) The Class A - Class B distinction could be extended or improved by re-defining the classes (on the basis of occupation for example), re-estimating the transition matrices conditional on class, or increasing the number of classes. Data to do this exists in the labor force survey, and occupation could be added to the state vector because it is carried in the SCF survey.
- (ii) The entire approach could be changed. One could imagine estimating distributions of weeks employed, weeks unemployed, etc., perhaps as functions of macro-economic conditions, occupation or class, age, sex, region, and so on. The problem here would be one of data, since the labor force survey is not directly amenable to such an approach.



- (iii) The present Markov-Chain model could be extended to a 2nd order model, and could perhaps be made conditional on occupation or class (See (1) above). The data for this exists in the labor force survey. The number of conditioning variables (occupation, age, sex, region, etc.) is of course limited by the size of the survey sample. The present model used one method to disaggregate transition matrices to additional conditioning variables, and further research could also be done in this area.

There are several extensions that appear to be possible in the Market Income Block.

- (i) The income state variables could be extended. It would be useful, for example, to distinguish various categories of employment income: self employment income (non-farm and farm), and wages and salaries. This would require a model in which the change in one of these income components was made conditional on all of the other components (including the relationship between employment and investment income). The present model for the most part assumes that the various income components are independent of one another. The DNR longitudinal data would be the only source of parameters, but there is sufficient data there to carry out the necessary estimations.



- (ii) The income change process itself could be modeled as a function of (perhaps) macro-parameters such as the rate of inflation, the rate of change of GNP, and so on. This would be in addition to the recognition that is already made of demographic variables and would introduce an implicit time variance to the transition matrices. Again, the DNR data would probably be sufficient here.
- (iii) Because of the obvious relationship between employment and income, it would be very desirable to eliminate the present general distinction between the Activity and Market Income Blocks. To some extent these blocks are at present tied together, of course, because it is necessary that the employment and income variables be consistent. One could improve on this, however, by thinking of activity-income as one distinct process, and conceiving of a model that would take account simultaneously of all of the activity-income variables. The difficulty here is one of data. Ideally, one would like to have a longitudinal data base that contained both income and activity variables. Unfortunately the two best data bases - the labor force survey and the DNR file - do not fulfill this requirement. The DNR data does not contain employment data, and the labor force survey does not contain income data. The closest approximation to the ideal is perhaps the UIC-DNR merge file. This data base suffers, unfortunately, from being out of date, and from various weaknesses





associated with the UIC data (the activity variables for the whole population - school, NLF, employment, unemployment - are not covered adequately). Perhaps one could construct a new data base, linking UIC records, DNR records, and some parts of the labor force survey. This would, of course, be a complex and costly process.

- (iv) The present model was estimated primarily from the UIC-DNR data base. Many parameters could perhaps be improved and conditioned on more demographic characteristics if the larger DNR data base were used. This of course would be much more expensive and much more time consuming (due to the difficulty of special-request access to the DNR data).









